Représentations robustes de documents bruités dans des espaces homogènes

Mohamed Morchid

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE
MINISTÈRE DE L'ENSEIGNEMENT
SUPÉRIEUR ET DE LA RECHERCHE

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 536 « Sciences et Agrosciences »
Laboratoire d'Informatique (EA 4128)

Représentations robustes de documents bruités dans des espaces homogènes

par
Mohamed Morchid

Soutenue publiquement le 25 Novembre 2014 devant un jury composé de :

M. Jérôme BELLEGARDA	Directeur, Apple Computer Inc., États-Unis	Rapporteur
M. Laurent BESACIER	Professeur, LIG, Grenoble, France	Rapporteur
M. François YVON	Professeur, Universirté Paris-Sud, France	Rapporteur
M. Youssef HAMADI	Sr. Researcher, Microsoft Research, Royaume-Uni	Examinateur
M. Frédéric BÉCHET	Professeur, LIF, Marseille, France	Examinateur
M. Benjamin PIWOWARSKI	Chargé de Recherche CNRS, LIP6, Paris, France	Examinateur
M. Georges LINARÈS	Professeur, LIA, Avignon, France	Directeur
M. Richard DUFOUR	Maître de Conférences, LIA, Avignon, France	Co-Encadrant

---

Laboratoire Informatique d'Avignon

Remerciements

Je tiens dans un premier temps à remercier Jérôme Bellegarda, Laurent Besacier ainsi que François Yvon pour avoir accepté de rapporter cette thèse. Je tiens également à remercier Frédéric Béchet, Benjamin Piwowarski ainsi que Youssef Hamadi pour avoir accepté d'être examinateurs de cette thèse.

Merci aussi à Mohamed B. et Driss pour tous ces bon moments passés au sein du LIA. J'ai également une pensée pour l'équipe administrative (Mireille, Dominique, Laurence, Simone, Cathy, Evelyne, Gisèle, ...) qui m'a apporté son soutien et surtout sa bonne humeur. Je tiens également à remercier l'équipe réseau et recherche opérationnelle pour avoir supporté mes visites journalières potentiellement perturbantes.

Lors de mon stage de 3 mois chez Microsoft Research Cambridge, j'ai été encadré par Yi et Youssef, et je vous remercie pour votre accueil, votre gentillesse, votre bonne humeur ainsi que pour votre apport scientifique durant toute cette période.

Tous les moments passés dans la cafétéria du LIA n'auraient pas été les mêmes sans les échanges parfois scientifiques et souvent humoristiques avec Rachid, Carole, Fabrice, Marc, Juan-Manuel, JP, Philou, Pierrot et d'autres. Merci également à toi Corinne pour ta bonne humeur ainsi que pour ta gentillesse. J'ai également apprécié ton amitié et surtout ta franchise Fabrice tout au long de ces années. Florent tu as été mon ami depuis mon arrivée en stage de Master 2 jusqu'à aujourd'hui et Waad qui a partagé mon bureau. Je vous remercie pour votre amitié sincère qui m'a bien aidé durant tout ce temps. Je remercie évidement Georges pour la confiance qu'il a eu en moi et pour son soutient scientifique ainsi que son amitié. J'ai également une tendre pensée pour Renato qui m'a soutenu durant ces années et m'a apporté beaucoup scientifiquement ainsi que personnellement. Merci également à Richard qui a été plus qu'un ami. Je remercie également profondément tous les personnes composant le LIA et le CERI que je ne cite pas ici mais qui resteront à jamais dans mes souvenirs.

Souvent on qualifie une personne réalisant le travail de thèse comme une personne courageuse. Le visage du véritable courage je l'ai rencontré il y a plus de dix ans. Merci de m'avoir transmis ta volonté, de m'avoir soutenu et merci pour ton amour Claire.

Je ne peux rédiger ces remerciements sans avoir un mot pour ma famille Ibrahim, Khadija, Bilal, Zohra, Abdelaziz ainsi que pour mes parents Belfakih et Laaidia.

Mes cheveux ressentent encore les caresses de ta douce main quand j'étais encore un enfant. Ton amour, ta tendresse ainsi que ton soutient sont comme ces caresses, à jamais présents dans ma mémoire ... alors merci pour tout maman ...

Résumé

En recherche d'information, les documents sont le plus souvent considérés comme des "sacs-de-mots". Ce modèle ne tient pas compte de la structure temporelle du document et est sensible aux bruits qui peuvent altérer la forme lexicale. Ces bruits peuvent être produits par différentes sources : forme peu contrôlée des messages des sites de *micro-blogging*, messages vocaux dont la transcription automatique contient des erreurs, variabilités lexicales et grammaticales dans les forums du Web... Le travail présenté dans cette thèse s'intéresse au problème de la représentation de documents issus de sources bruitées.

La thèse comporte trois parties dans lesquelles différentes représentations des contenus sont proposées. La première partie compare une représentation classique utilisant la fréquence des mots à une représentation de haut-niveau s'appuyant sur un espace de thèmes. Cette abstraction du contenu permet de limiter l'altération de la forme de surface du document bruité en le représentant par un ensemble de caractéristiques de haut-niveau. Nos expériences confirment que cette projection dans un espace de thèmes permet d'améliorer les résultats obtenus sur diverses tâches de recherche d'information en comparaison d'une représentation plus classique utilisant la fréquence des mots. Le problème majeur d'une telle représentation est qu'elle est fondée sur un espace de thèmes dont les paramètres sont choisis empiriquement.

La deuxième partie décrit une nouvelle représentation s'appuyant sur des espaces multiples et permettant de résoudre trois problèmes majeurs : la proximité des sujets traités dans le document, le choix difficile des paramètres du modèle de thèmes ainsi que la robustesse de la représentation. Partant de l'idée qu'une seule représentation des contenus ne peut pas capturer l'ensemble des informations utiles, nous proposons d'augmenter le nombre de vues sur un même document. Cette multiplication des vues permet de générer des observations "artificielles" qui contiennent des fragments de l'information utile. Une première expérience a validé cette approche multi-vues de la représentation de textes bruités. Elle a cependant l'inconvénient d'être très volumineuse, redondante, et de contenir une variabilité additionnelle liée à la diversité des vues.

Dans un deuxième temps, nous proposons une méthode s'appuyant sur l'analyse factorielle pour fusionner les vues multiples et obtenir une nouvelle représentation robuste, de dimension réduite, ne contenant que la partie "utile" du document tout en réduisant les variabilités "parasites". Lors d'une tâche de catégorisation de conversations, ce processus de compression a confirmé qu'il permettait d'augmenter la robustesse de

la représentation du document bruité.

Cependant, lors de l'élaboration des espaces de thèmes, le document reste considéré comme un "sac-de-mots" alors que plusieurs études montrent que la position d'un terme au sein du document est importante. Une représentation tenant compte de cette structure temporelle du document est proposée dans la troisième partie. Cette représentation s'appuie sur les nombres hyper-complexes de dimension 4 appelés *quaternions*. Nos expériences menées sur une tâche de catégorisation ont montré l'efficacité de cette méthode comparativement aux représentations classiques en "sacs-de-mots".

Mots clés : *Représentation robuste, document bruité, allocation latente de Dirichlet, multi-vues, analyse factorielle, quaternion.*

Abstract

In the Information Retrieval field, documents are usually considered as a "bag-of-words". This model does not take into account the temporal structure of the document and is sensitive to noises which can alter its lexical form. These noises can be produced by different sources : uncontrolled form of documents in micro-blogging platforms, automatic transcription of speech documents which are error-prone, lexical and grammatical variabilities in Web forums... The work presented in this thesis addresses issues related to document representations from noisy sources.

The thesis consists of three parts in which different representations of content are available. The first one compares a classical representation based on a term-frequency representation to a higher level representation based on a topic space. The abstraction of the document content allows us to limit the alteration of the noisy document by representing its content with a set of high-level features. Our experiments confirm that mapping a noisy document into a topic space allows us to improve the results obtained during different information retrieval tasks compared to a classical approach based on term frequency. The major problem with such a high-level representation is that it is based on a space theme whose parameters are chosen empirically.

The second part presents a novel representation based on multiple topic spaces that allow us to solve three main problems : the closeness of the subjects discussed in the document, the tricky choice of the "right" values of the topic space parameters and the robustness of the topic-based representation. Based on the idea that a single representation of the contents cannot capture all the relevant information, we propose to increase the number of views on a single document. This multiplication of views generates "artificial" observations that contain fragments of useful information. The first experiment validated the multi-view approach to represent noisy texts. However, it has the disadvantage of being very large and redundant and of containing additional variability associated with the diversity of views. In the second step, we propose a method based on factor analysis to compact the different views and to obtain a new robust representation of low dimension which contains only the informative part of the document while the noisy variabilities are compensated. During a dialogue classification task, the compression process confirmed that this compact representation allows us to improve the robustness of noisy document representation.

Nonetheless, during the learning process of topic spaces, the document is considered as a "bag-of-words" while many studies have showed that the word position in a

document is useful. A representation which takes into account the temporal structure of the document based on hyper-complex numbers is proposed in the third part. This representation is based on the hyper-complex numbers of dimension four named *quaternions*. Our experiments on a classification task have showed the effectiveness of the proposed approach compared to a conventional "bag-of-words" representation.

Keywords : *Robust representation, noisy document, latent Dirichlet allocation, multi-views, factor analysis, quaternion.*

Table des matières

Introduction

Enjeux liés au traitement de l'information

Depuis l'avènement, ces dernières décennies, de nouvelles super-puissances considérées jusqu'alors comme des pays en voie de développement, tels la Chine ou certains pays d'Amérique du Sud, les relations de subordination liant les entités de toutes tailles se sont déplacées du terrain économique, militaire ou politique, vers une relation dictée par l'acquisition d'une information stratégique, précise, pertinente et fiable (Niquet, 2000; infoGuerre, 2000; Susbielle, 2006).

La bataille que se livrent les institutions pour se munir d'outils efficaces, leur permettant de disposer d'une information fiable, les pousse à mobiliser toujours plus de ressources financières, de personnels et de temps de développement (Susbielle, 2006; Assange, 2013).

Notre société est donc entrée dans un nouvel âge, celui de l'acquisition et du traitement de l'information sur des médias, avec des technologies d'abord rudimentaires puis plus sophistiquées, destinées à collecter, sélectionner ou représenter l'information.

La demande qualitative et quantitative d'une information pertinente a entraîné une croissance de l'intérêt porté aux sciences liées à l'acquisition et au traitement des données. Les enjeux économiques, politiques et sociétaux, pour les institutions ou entreprises, sont considérables dans une société ultra-connectée et désireuse de partager toujours plus d'informations.

L'information est essentiellement véhiculée à travers des médias numériques prenant des formes diverses au fur et à mesure que les technologies propres à ces supports se développent et se complexifient. Ces technologies offrent des appareils toujours plus rapides, constamment connectés et à la capacité de stockage croissante. Ainsi, à la complexité liée à la diversité des supports et de l'information, s'est ajoutée une complexité relative à la capacité de stockage qui croît de manière exponentielle. En effet, le nombre de documents disponibles dans les bases de données et le nombre grandissant d'informations qu'elles soient textuelles, visuelles, ou auditives, contribuent à la difficulté de la tâche de traitement de très gros volumes d'informations hétérogènes.

De nouveaux domaines de recherche ont vu le jour, portés par cette nécessité de traiter ces ensembles de documents de dimension très importante, d'une manière efficace

et rapide. Ces nouveaux pans de la recherche vont de l'acquisition de connaissances dans des ensembles d'informations de grandes tailles (Fouille de textes ou *Text mining*), en passant par la recherche d'une information pertinente au sein d'un corpus de documents (Recherche d'Information (RI) ou *Information Retrieval* (IR)), pour finir par le domaine, plus vaste, du traitement du langage naturel (Traitement Automatique de la Langue (TAL) ou *Natural Language Processing* (NLP)) ou de données multimédias. Les parties suivantes reviennent sur les évolutions de chacun de ces domaines de recherche.

Recherche d'information

La RI "**étudie la manière de retrouver des informations dans un corpus. Celui-ci est composé de documents d'une ou plusieurs bases de données qui sont décrits par un contenu ou les méta-données associées. Les bases de données peuvent être relationnelles ou non structurées, telles celles mises en réseau par des liens hypertextes comme dans le World Wide Web (WWW), l'Internet et les Intranets. Le contenu des documents peut être du texte, des sons, des images, ou des données.**" [1]

Il s'agit donc d'extraire de l'information pertinente sachant une requête délivrée par un utilisateur au sein d'un corpus de documents textuels, d'images, ou de sons. Cet objectif assez général peut s'appliquer à des tâches comme la tâche de Question-Réponse (QR) ou *Question-Answering* (QA), l'analyse des contenus dans les réseaux sociaux, la recommandation, le résumé automatique, ...

Le domaine de la RI est confronté à différentes questions de fond :

- Comment trouver des documents similaires à un document donné ?
- Sachant la taille des ensembles de données contenant l'information recherchée, comment traiter cette masse d'information dans un délai que l'utilisateur imagine instantané ?
- Quelle est la manière la plus pertinente de présenter les résultats ainsi trouvés ?
- Quelle est la meilleure correspondance entre l'idée véhiculée dans la requête d'un utilisateur et celle contenue dans les documents composant le corpus de recherche ?

Nous constatons que les objectifs visés sont souvent contradictoires (vitesse contre précision, exhaustivité contre pertinence, ...).

La recherche d'information trouve ses fondements durant les années 50 lors de l'avènement des premiers ordinateurs personnels (*Personal Computer* (PC)) (Stock, 2007). Le domaine à cependant pris une autre dimension avec le développement des réseaux. Dans la vie quotidienne des utilisateurs de terminaux connectés à Internet tels

1. **http://fr.wikipedia.org/wiki/Recherche_d%27information**

les ordinateurs, les tablettes, ou les téléphones intelligents (*smartphones*), la recherche d'information dans des bases de données de grandes tailles se fait via des moteurs de recherche qui sont devenus des points d'entrée incontournables de l'Internet. Selon (Baeza-Yates et al., 1999), la RI peut être partagée en trois tâches principales : les tâches de recherche d'information dites *ad-hoc*, l'extraction de documents pertinents sachant des caractéristiques de l'utilisateur et la navigation.

La tâche d'extraction de documents sachant une requête délivrée par un utilisateur, via un moteur de recherche sur Internet par exemple, peut se révéler excessivement coûteuse en temps d'exécution si elle est réalisée de manière directe sans traitement préalable. Dans (Manning et al., 2008), les auteurs proposent de trouver les documents correspondant le mieux à une requête donnée en parcourant de manière linéaire l'ensemble des documents composant le corpus. Ensuite, ils cherchent à extraire les passages dans les documents contenant les termes de la requête. Il est donc nécessaire, pour chacune des requêtes reçues, de parcourir à nouveau l'ensemble du corpus, ce qui rend cette approche très coûteuse en temps de traitement, surtout lorsqu'il s'agit d'une base de connaissance telle qu'Internet.

Afin d'éviter ce traitement fastidieux à chaque nouvelle requête, il apparaît nécessaire de procéder à un pré-traitement consistant à stocker l'information liant les occurrences de chacun des termes du vocabulaire avec les documents composant le corpus. Cette structure de données est communément appelée "index inversé" du corpus de documents (Weiss et al., 2010). Un index inversé est un dictionnaire comportant, pour chacun des termes le composant, son emplacement dans un corpus (documents contenant au moins une occurrence de ce terme) et éventuellement d'autres informations caractérisant le terme, comme par exemple la fréquence inverse dans le corpus (*Inverse Document Frequency* (IDF)) (voir annexe A).

La figure 1 illustre les différentes étapes du calcul d'un index inversé depuis un corpus de documents. La première phase consiste à découper chacun des documents en une succession de termes suivant le processus de **grep***ping* ("**g**ets **r**egular **e**x**p**ressions") afin d'obtenir une première représentation du corpus (figure 1-index). Suite à cette première étape, une phase de réorganisation des termes du vocabulaire selon l'ordre alphabétique (figure 1-index ordonné), puis un décompte du nombre d'occurrences de chacun des termes, permettent d'aboutir à l'index inversé du corpus de documents (figure 1-index inversé).

Cette représentation du corpus sous forme d'un index inversé permet de traiter les données initialement représentées dans un espace discret (terme et document) comme une matrice de données réelles (occurrences de mots au sein du corpus), ou sous forme d'une matrice de valeurs dites "flottantes" (nombre d'occurrences des termes divisé par le nombre total de termes composant le corpus). Elle permet également un gain de temps en séparant la fréquence des termes de leur position dans le corpus (figure 1-Id.Doc.).

Les méthodes de recherche d'information s'appuyant sur cette nouvelle représentation du corpus de documents comme une matrice (A) sont détaillées dans le chapitre 2.

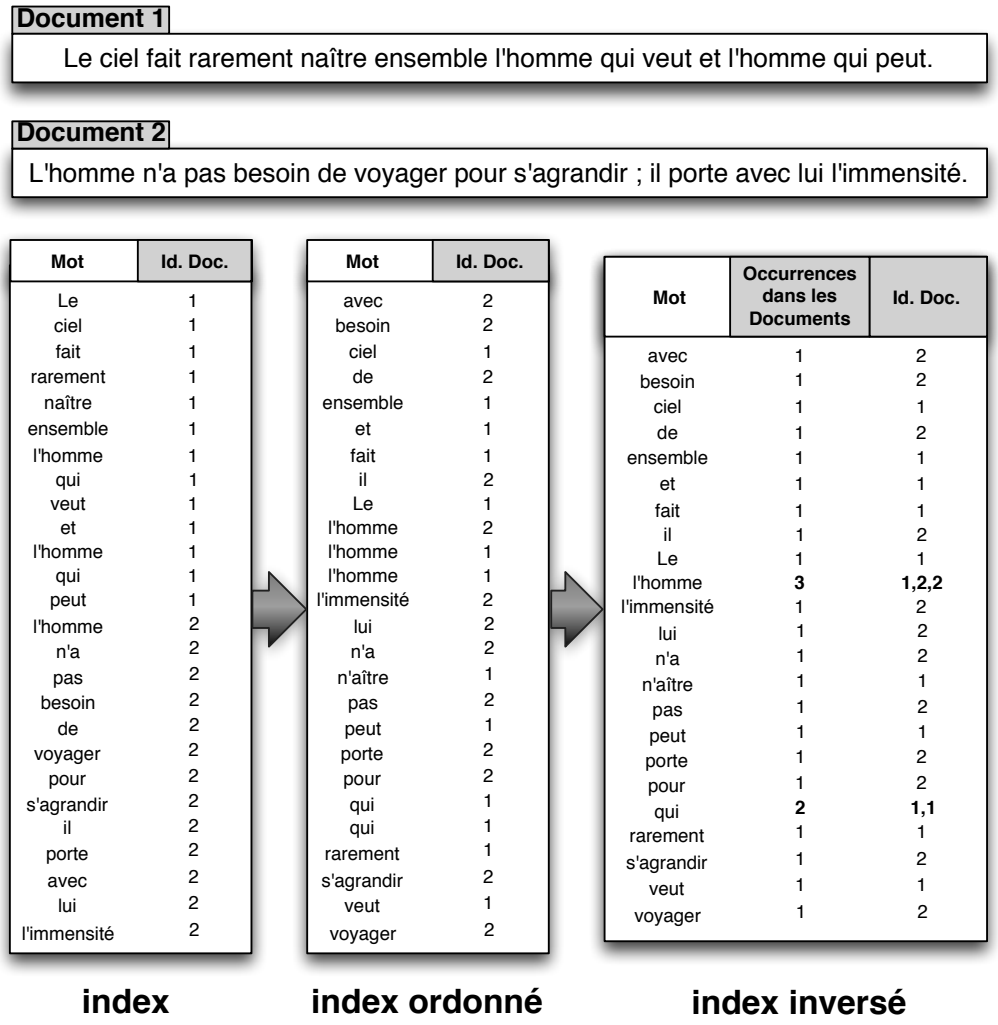


FIGURE 1 – *Exemple adapté de (Manning et al., 2008) d'un index inversé constitué depuis un corpus de deux documents (deux citations de François René de Chateaubriand, Extraits de "Mémoires d'outre-tombe" pour le document 1 et de "De la restauration et la monarchie élective" pour le document 2).*

Traitement Automatique du Langage Naturel

La recherche d'information (RI) a pour objectif d'extraire un ensemble de documents pertinents sachant une requête délivrée par l'utilisateur, en général à un moteur de recherche. Le domaine du Traitement Automatique du Langage Naturel (TALN) consiste, lui, au traitement du langage dit "humain". Ce domaine est issu de l'Intelligence Artificielle (IA) et est connexe à de nombreux autres domaines proches de l'IA, comme le traitement de la parole ou l'indexation de documents multimédias.

Les outils développés par les scientifiques du TALN sont fréquemment utilisés

dans le domaine de la RI (Stock, 2007). Ces pré-traitements permettent une exploitation efficace des corpus de très grandes tailles, comme par exemple, les données issues d'Internet indexées par les moteurs de recherche (Kruschwitz, 2005). Parmi ces outils performants issus du TALN et largement utilisés dans le domaine de la RI, nous trouvons :

- les procédés de segmentation d'un document textuel en une succession de termes (*tokens*) (Manning et Schütze, 1999) de phrases, de thèmes...
- l'utilisation d'une liste de mots au contenu informatif faible, pour ne garder que les termes les plus pertinents du document (Manning et Schütze, 1999),
- les modules de reconnaissance d'entités nommées (Stock, 2007),
- les outils d'étiquetage (*tagging*) appelés *Part-Of-Speech (POS) tagging* permettant de définir la classe morpho-synthaxique d'un terme au sein d'une phrase (verbe, nom, préposition, ...) (Manning et Schütze, 1999),
- les systèmes de résolution d'anaphores (les pronoms comme "elle" ou "lui") ainsi que les ellipses (omission dans une phrase d'un terme) (Stock, 2007).

Fouille de textes

La fouille de textes s'intéresse à l'analyse et à l'interprétation de documents textuels. C'est un domaine connexe à la RI, au TALN (Hearst, 1999) et au traitement statistique des données. Ce domaine se décline en un ensemble de tâches devenues aujourd'hui essentielles pour un traitement efficace des bases de connaissance disponibles notamment sur Internet et dans les entreprises. Nous retrouvons ainsi les tâches :

- de résumé automatique de documents textuels ou parlés (Fan et al., 2006),
- d'extraction de mots-clés (Berry et Kogan, 2010),
- de recherche de concepts dans des bases de documents (Blei et Lafferty, 2006b; Fan et al., 2006; Wang et McCallum, 2006; Wei et Croft, 2007),
- de catégorisation et classification de documents (Feinerer, 2008; Berry et Kogan, 2010).

Cette dernière tâche de catégorisation de documents sera largement traitée dans la suite de ce manuscrit.

Problématiques

Un des problèmes centraux en fouille de textes est la catégorisation de documents textuels ou parlés. Ce problème a été largement traité dans la littérature, dans des contextes applicatifs assez variés, sur des données écrites ou de la parole transcrite. Les systèmes à l'état-de-l'art obtiennent de bon résultats sur des données "propres" (documents audio enregistrés dans des conditions optimales, documents textuels "bien" écrits, respectant les règles orthographiques et grammaticales, ...).

Malheureusement, les données d'Internet sont très variables et les contextes de production et d'acquisition permettent rarement de réunir les conditions optimales nécessaires pour atteindre ce bon niveau de performance. Le domaine de la catégorisation de documents se heurte donc à une difficulté liée au développement de ces grandes masses d'information hétérogènes.

Ces particularités rendent difficiles l'archivage structuré des documents, qui est souvent considéré comme une problématique cruciale pour le bon fonctionnement ou le développement des entreprises. Elles limitent aussi l'accès à la masse d'information disponible sur un réseau de plusieurs milliards d'individus.

La catégorisation, et, plus généralement l'analyse des contenus dans des collections de taille aussi importante, se fonde sur un ensemble de descripteurs qui doivent caractériser au mieux le contenu d'un document. Cet ensemble de descripteurs doit être le plus réduit possible, pour permettre des analyses robustes et rapides, tout en préservant l'information utile contenue dans le document. Il doit aussi préserver les relations statistiques des documents entre eux. Un point absolument critique pour l'analyse des contenus est l'espace dans lequel les documents sont codés et analysés.

Les techniques de représentation dans un espace de dimension réduite ont fortement évolué ces deux dernières décennies. Nous sommes passés d'une représentation par de simples vecteurs d'occurrences de mots à des représentations plus abstraites, qui cherchent à tenir compte de la structure des contenus sémantiques du document. Ces techniques ont néanmoins été principalement développées pour des données textuelles propres, générées dans des environnements contrôlés.

La difficulté majeure qui se pose alors pour appliquer ces techniques à des données fortement bruitées est le choix d'un nombre réduit de descripteurs robustes permettant une bonne représentation du document en dépit de la variabilité des contenus, des contextes de production et de diffusion, des supports, ...

Les travaux réalisés dans cette thèse se placent dans le contexte de documents bruités issus de transcriptions automatiques d'un système de reconnaissance automatique de la parole ou provenant de messages partagés par les usagers du Web. Ces documents contiennent une partie considérée comme utile et une autre considérée comme « nuisible », que nous appellerons "bruit". Ce bruit peut prendre des formes très diverses, notamment :

— de messages vocaux mal transcrits par le système de reconnaissance automatique de la parole,

— de messages écrits dont le vocabulaire ou le style d'écriture est atypique selon les habitudes des internautes (Twitter, messagerie instantanée, ...).

L'impact du bruit sur la capacité des systèmes à analyser les contenus est intrinsèquement lié à la représentation des documents. Par exemple, le bruit peut se manifester par des mots qui sont mal transcrits ou mal écrits, risquant d'altérer les représentations fondées sur la distribution des termes au sein du document. La recherche d'une représentation permettant de tolérer les bruits, qu'il s'agisse d'erreurs de transcription ou d'une autre variabilité de forme des documents, est un enjeu majeur dans le domaine du traitement de la langue. Les travaux entrepris lors de cette thèse cherchent, dans un premier temps, à évaluer différentes représentations de documents considérés comme bruités.

La représentation la plus classique d'un document s'appuie sur la fréquence des termes le composant. Il est clair que ces descripteurs sont fortement dépendants des occurrences éventuellement erronées des termes. Des modèles plus élaborés ont été proposés ensuite pour obtenir une représentation plus abstraite du document s'appuyant sur des espaces thématiques. Ces techniques reposent sur l'idée que projeter la forme de surface d'un document dans un espace de plus haut niveau devrait permettre de limiter le bruit observé au niveau lexical. Ces deux représentations en vecteurs de fréquences et représentation thématique, sont évaluées lors de la première étude pour extraire un ensemble de descripteurs pertinents, limitant le bruit contenu dans les documents.

Deux problèmes majeurs lors de l'élaboration de ces espaces abstraits sont ensuite traités : le choix de la configuration des paramètres de tels modèles ainsi que la proximité entre les sujets traités dans ces documents. En effet, lors d'une tâche de catégorisation de documents, les classes ou sujets, sont souvent très proches. Pour pallier cette difficulté, une étude sur une représentation multi-vues simple d'un document est proposée pour tenir compte des différents aspects du document. Cette représentation multiple simple permet d'obtenir de bon résultats, mais ne permet pas de résoudre la problématique liée au choix des paramètres du modèle thématique.

Par la suite, une nouvelle représentation s'appuyant sur des espaces thématiques multiples est proposée afin de répondre au difficile choix des hyper-paramètres du modèle. Cette technique multiplie, dans un premier temps, les espaces de représentation en faisant varier les hyper-paramètres. Ces représentations complémentaires sont ensuite compactées pour obtenir finalement, un espace réduit permettant une catégorisation robuste, stable et efficace de documents bruités.

Ces représentations proposent donc d'améliorer la robustesse en se plaçant au niveau sémantique et en multipliant les vues sur les documents. Cependant, ces représentations ne tiennent pas compte de la structure temporelle du document qu'elles considèrent comme un "sac de mots". Pourtant, plusieurs études montrent que la position d'un terme au sein du document est une donnée importante dont il faudrait tenir compte.

Cette importance de la structure temporelle est bien illustrée par des expériences réalisées dans le domaine du résumé automatique. Cette tâche a pour objectif de réduire la taille d'un document en ne conservant que l'information essentielle. Les sys-

tèmes performants surpassent difficilement une technique très simple qui consiste à ne conserver que les premières phrases contenues dans un document.

Une dernière contribution est d'intégrer l'information temporelle dans les espaces de représentation. Cette contribution repose sur un codage des distributions lexicales dans des nombres hyper-complexes.

Ce manuscrit traite donc de la représentation de documents bruités pour l'analyse des contenus. Les principales questions auxquelles nous chercherons à apporter des éléments de réponse sont :

- À quel niveau l'information doit-elle être représentée ?
- Quels types d'information doivent être intégrées à la représentation ?
- Quels formalismes utiliser pour représenter ces informations ?

Plan du manuscrit

Toutes ces questions sont traitées dans les chapitres suivants de ce manuscrit afin d'y apporter des éléments de réponses et, peut être, de nouvelles interrogations. Dans un premier temps, un état de l'art est présenté dans le chapitre 1 sur la représentation des documents dans des espaces vectoriels continus. Ce chapitre retrace l'historique des représentations vectorielles depuis la plus élémentaire, fondée sur la fréquence de mots, à d'autres plus sophistiquées, utilisant des dimensions abstraites considérées comme des concepts.

Le chapitre 2 décrit, dans un premier temps, les différentes méthodes de représentation d'un document et introduit la notion de représentation sémantique. Ce chapitre met en évidence l'apport d'une représentation de haut-niveau d'un document bruité ou de taille réduite, dans la tâche de catégorisation de documents. Cette représentation, dans un espace de thèmes unique d'un document bruité appliquée à des tâches de recherche d'information, permet de replacer le document dans un contexte précis et d'enrichir la représentation.

Le chapitre 3 présente différentes études évaluant l'apport d'une représentation dans de multiples espaces de thèmes d'un document bruité issu du Web ou de transcriptions automatiques lors de tâches de classification ou de catégorisation. Les conclusions de ce chapitre permettent d'entrevoir de nouvelles possibilités liées à la représentation multiple de documents. Ces représentations multiples introduisent néanmoins une information résiduelle lorsque les documents sont projetés dans des espaces multiples de thèmes. En effet, l'information utile est entremêlée à une part de variabilité "nuisible" liée aux différentes projections ainsi qu'aux erreurs du système de transcription automatique. Les chercheurs du domaine de la vérification du locuteur ont déjà fait

face a ce problème et ont pu (en partie) le résoudre en utilisant de nouvelles approches issues de l'analyse factorielle.

L'application de ces méthodes aux documents textuels bruités est détaillée dans le chapitre 4. Les résultats obtenus permettent de penser que la variabilité nuisible peut être efficacement compensée dans le cas de documents textuels fortement bruités. Les résultats des chapitres 2 à 4 sont obtenus avec une représentation thématique du document, ne prenant pas en compte la position des termes au sein de ce document. Ceci est particulièrement handicapant lorsque ces documents contiennent plusieurs sujets et ont une structuration temporelle forte.

Le chapitre 5 montre l'apport d'une représentation du document codant les relations entre les différentes occurrences des mots au sein du document via une représentation dans l'espace des hyper-complexes, appelée *quaternion*. Les conclusions sur les différentes représentations de documents bruités ainsi que des perspectives de recherche sont présentées en fin de ce manuscrit. La partie IV récapitule les différents travaux entrepris pour une représentation robuste dans les chapitres précédents, et propose un ensemble de perspectives pour une représentation robuste de documents bruités.

Première partie

Projection de documents bruités dans un espace thématique

Chapitre 1

État de l'art de la représentation de documents dans des espaces de thèmes

Sommaire

Résumé

Ce chapitre présente les modèles thématiques pour la représentation de textes. Nous décrivons d'abord assez rapidement les principaux paradigmes de représentation de texte proposés dans le passé puis nous développons plus longuement les modèles à base d'analyse sémantique latente, sur lesquels reposent une bonne partie des contributions décrites dans les chapitres suivants.

1.1 Introduction

Les travaux réalisés dans cette thèse ont pour objectif de proposer des solutions nouvelles permettant de représenter efficacement des documents fortement bruités lors de tâches de traitement de l'information. Plusieurs modélisations du document ont été proposées depuis les années 70 avec les travaux proposés par l'équipe du chercheur Gerard Salton (Salton, 1971). Les premiers travaux présentés ont permis de proposer les premières représentations vectorielles d'un document considéré jusqu'alors comme une simple succession de termes. Depuis, des représentations plus évoluées ont été proposées.

La représentation d'un texte repose sur la représentation des mots qui le compose, mais ne peut pas être réduite à cette simple représentation : elle doit coder la structure globale du document. Plusieurs représentations des termes composant le document ont été proposées. La plus classique s'appuie sur la fréquence des termes et est présentée dans la section 1.2. La représentation d'un terme peut également s'appuyer sur son contexte d'apparition, par exemple en modélisant le lien entre le mot et sa classe grammaticale (WordNet (Miller, 1995)).

Récemment, (Bengio, 2009; Collobert et al., 2011; Deng et al., 2013; Heck et al., 2000; Tur et al., 2012) ont appliqué avec succès, à la recherche d'information, une variété d'applications s'appuyant sur les réseaux de neurones. Ces méthodes ont permis de nouvelles représentations des termes appelées *Word embeddings*. Ce modèle est devenu un champ de recherche considérable dans le domaine du traitement de la langue. La représentation des séquences de mots par les *word embeddings* est obtenue à l'aide de réseaux de neurones et introduite pour la première fois par (Hinton, 1986). Plus tard, les auteurs dans (Bengio et al., 2006) ont repris ces travaux pour proposer un modèle de langage fondé sur des probabilités de n-grammes apprises à l'aide de réseaux de neurones appelés réseaux de neurones profonds ou *Deep Belief Networks* (DBN) (Sarikaya et al., 2011).Dans (Mikolov et al., 2013), les auteurs proposent d'utiliser ces réseaux de neurones pour mettre en évidence les relations existantes entre les termes. Ces relations sont exprimées en fonctions des relations entre leurs vecteurs de représentation.

Les auteurs dans (Tur et al., 2012) proposent d'utiliser un réseau de neurones profond convexe (*Deep Convex Network* ou DCN (Deng et Yu, 2011)) en lieu et place du réseau DBN pour extraire, depuis des dialogues entre un humain et une machine, un ensemble de domaines et d'intentions de l'utilisateur pour parvenir à un objectif ou à une action. Ceci est dû à la difficulté de mise à l'échelle du DBN. De plus, les DCN obtiennent de meilleurs résultats en termes de précision (Deng et Yu, 2011). Les expérimentations sont menées dans une tâche de catégorisation de dialogues parmi 25 thèmes. Le système proposé s'appuyant sur un DCN est comparé à une méthode plus classique appelée le *Boosting* (Schapire et Singer, 2000). Les résultats observés montrent deux choses : le système fondé sur le DCN améliore la précision de la catégorisation comparativement à la méthode de *Boosting* ; il est aussi notable que les résultats obtenus sont équivalents avec l'utilisation des données annotées ou avec avec l'ensemble des données disponibles. Ces méthodes se concentrent sur la représentation des mots

plutôt que sur celle des documents, la représentation des documents étant le problème de fond traité dans les travaux présentés dans ce mémoire.

Plusieurs représentations du document ont été proposées dans le passé. La plus classique s'appuie sur un vecteur de fréquences de mots. Chacune des caractéristiques de ce vecteur contient la fréquence d'un terme du vocabulaire dans ce document. Ces représentations classiques du document s'appuient sur des "sacs-de-mots " (Salton, 1971) ou, dans la même idée, sur des "sacs-de-phrases" (Gruber et al., 2007; Nathan, 2009). La partie 1.2 présente ce modèle plus en détail. Une représentation plus évoluée s'appuyant sur le regroupement des mots en concepts est proposée dans (Sahlgren et Cöster, 2004). Les auteurs proposent d'utiliser une méthode fondée sur l'indexation aléatoire (*Random Initialization* (Kanerva et al., 2000)) pour regrouper les termes entre eux. Cette méthode s'appuie sur les cooccurrences des termes au sein du document. Ce processus est moins couteux en termes de temps de traitement que des approches comme la décomposition en valeurs singulière (SVD).

Certaines extensions de ce modèle ont été proposées par la suite. Dans (Piwowarski, 2013), l'auteur propose d'utiliser les probabilités quantiques pour représenter un document. Ces méthodes sont inspirées des formalismes probabilistes de la physique quantique proposés par (Van Rijsbergen, 2004). Ces représentations ont montré leur efficacité dans des tâches comme la tâche de question-réponse ou le résumé automatique.

Les contributions qui seront décrites tout au long de ce mémoire portent sur des représentations fondées sur la fréquence de mots et sur des espaces de thèmes essentiellement. Pour cette raison, ce chapitre se concentre principalement sur l'état-de-l'art les modèles à base de fréquence de mots ainsi que les représentations thématiques.

1.2 Modèle à base de fréquence de mots

Le domaine de la recherche d'information (RI), porté par le besoin de méthodes de traitement et d'accès à de grandes masses d'information (catégorisations, résumé automatique, extraction d'information, ... (Baeza-Yates et al., 1999)), a réalisé des progrès majeurs ces dernières années.

Une des premières méthodes proposées par les spécialistes de la RI est relativement simple, mais a largement été utilisée dans les moteurs de recherche avec une évidente efficacité (Salton, 1971). Cette méthode consiste à transformer un document contenu dans un corpus en un vecteur composé de la fréquence des mots contenus dans le vocabulaire. Plus communément appelée TF-IDF, cette méthode sélectionne un vocabulaire ("sac-de-mots" (Salton, 1989b)), et, pour chacun des mots contenus dans ce vocabulaire, elle détermine son nombre d'occurrences au sein du document (TF). Cette valeur, une fois normalisée, est comparée à la fréquence inverse du mot appelée *idf* (Jones, 1972). L'IDF mesure la rareté d'un mot au sein de l'ensemble des documents du corpus. Cette caractéristique est donc un bon indicateur du pouvoir discriminant du mot. Une discussion autour de l'intérêt du modèle et de ses dérivées est proposée dans (Claveau, 2012) (cf. annexe A).

L'utilisation du TF-IDF pour organiser une collection de documents a été massivement étudiée, tout particulièrement par G. Salton et al., pour une représentation vectorielle d'un document (Salton, 1971). Par la suite, les auteurs dans (Salton et Yang, 1973) montrent que le pouvoir discriminant d'un mot est proportionnel à la valeur du TF-IDF qui lui est associée.

La suite des travaux de l'équipe de G. Salton sur le pouvoir discriminant des mots au sein d'une collection de documents (Salton et al., 1975) les a conduit à proposer une méthode de quantification du pouvoir discriminant d'un terme. Considérant une collection de documents, dont chacun d'entre eux est représenté par un vecteur ayant pour dimension la taille du vocabulaire, il est possible de définir une mesure de "proximité" entre deux documents en calculant leur similarité (norme euclidienne ou cosinus par exemple). Deux documents seront considérés comme proches dans l'espace *terme-document* si la mesure de similarité entre ces deux documents est élevée et, *a contrario*, ces documents seront considérés comme partageant peu d'information si la distance qui les sépare est faible. Un terme est alors défini comme ayant un *fort* pouvoir discriminant si le fait d'apparaître dans un document permet de diminuer sa similarité avec les autres documents appartenant à la collection. Inversement, un terme a un pouvoir discriminant *faible* s'il "rapproche" le document dans lequel il apparaît des autres documents du corpus.

Dans (Salton et al., 1975), les auteurs définissent l'appartenance d'un terme discriminant à un document comme une augmentation de la densité de l'espace de représentation des documents (espace *terme-document*), en augmentant la distance entre le document dans lequel le mot apparaît et les autres. Inversement, un terme peu discriminant aura tendance à diminuer la densité de l'espace *document-document*. La valeur discriminante d'un mot est alors déterminée en calculant la densité de l'espace de représentation avant et après affectation de chaque terme à un document. Les termes formant le vocabulaire sont ensuite classés selon la valeur de leur TF-IDF. Trois catégories de mots discriminants sont définies dans (Salton et al., 1975) : *fort*, *faible* et *neutre*. Il est à noter que les auteurs sont arrivés à la conclusion que la majorité des termes contenus dans une collection de documents possède un pouvoir discriminant *faible*. Ils ont ensuite cherché à appliquer cette méthode à d'autres tâches de la recherche d'information (Salton et McGill, 1983; Salton et Buckley, 1988).

La représentation d'une collection de documents dans un espace thématique a été utilisée comme une approche à l'état-de-l'art, puis comme une méthode de référence permettant des évaluations comparatives avec d'autres méthodes plus élaborées. Dans (Cohen, 1996), les auteurs utilisent une pondération TF-IDF et un système de recherche de mots-clés à base de règles, nommé RIPPER (William et al., 1995), pour classifier un ensemble de courriels. Les deux méthodes obtiennent des résultats équivalents en termes de généralisation ou de classification. L'approche TF-IDF a essentiellement été utilisée dans la recherche de mots discriminants et dans les tâches de *question-réponse*. Ce modèle était considéré comme la référence lors de l'avènement des

moteurs de recherche tels que Lemur [1].

Une représentation s'appuyant uniquement sur la fréquence des termes composant un document donne peu (ou pas) d'information sur la structure statistique inter- ou intra-documents. Les modèles thématiques, détaillés dans la section suivante, permettent de résoudre ce problème. En effet, nous verrons que lors de la phase d'estimation, les relations entre les documents sont codées par les variables latentes composant le modèle.

1.3 Modèles thématiques

Les modèles fondés sur la fréquence des termes permettent de représenter les documents dans un espace commun. Cependant, la dimension de cet espace est très élevée car cette dimension est liée (proportionnelle) à la taille du vocabulaire. Les représentations dans cet espace sont très creuses. Une autre faiblesse de ce modèle est que les dépendances statistiques inter- et intra-documents sont très faiblement prises en compte. Un espace de thèmes ou *topic model* (TM) permet de définir cette structure liant les termes composant les documents, ainsi que les documents entre eux. Le nom de "thème" représente la variable modélisant la relation, ou distribution cachée devant être estimée, liant les mots composant un vocabulaire et leurs occurrences au sein des documents composant le corpus d'apprentissage. Les espaces de thèmes reposent sur l'idée que la structure sémantique des documents est latente, et qu'elle peut être extraite par l'analyse, à grande échelle, des distributions lexicales dans des corpus de grandes tailles. Cette hypothèse s'appuie sur le fait qu'un document puisse être généré par un processus stochastique qui est ensuite lui-même "renversé" (Blei et McAuliffe, 2007) par les méthodes de *machine learning*. Ces méthodes fournissent une estimation des variables latentes liant le vocabulaire et la distribution des mots au sein d'un corpus. Ces variables latentes, composant les différents modèles, permettent l'utilisation de corpus de grandes tailles comme des sources d'information *a priori* qui peuvent être utilisées dans des tâches de RI.

1.3.1 Analyse sémantique latente (LSA)

Les chercheurs en RI ont donc proposé de nouvelles méthodes de réduction de l'espace de représentation permettant automatiquement de trouver des relations sémantiques entre les mots et les documents formant le corpus ou la collection de documents. La première est l'indexation sémantique latente ou analyse sémantique latente (*Latent semantic indexing or analysis* (Analyse Sémantique Latente ou *Latent Semantic Analysis* (LSA)/LSA)) (Bellegarda, 1997; Deerwester et al., 1990; Bellegarda, 2000). LSA est un

1. Lemur a été développé à l'University de Massachusetts à Amherst (**http://www.lemurproject.org/**), Terrier à l'University of Glasgow (**http://ir.dcs.gla.ac.uk/terrier/**) et le système MG à l'University of Melbourne (**http://www.cs.mu.oz.au/mg/**). Lucene a été conçu par D. Cutting (**http://lucene.apache.org/**).

paradigme original formulé dans le contexte de la RI et très souvent utilisé, en particulier durant la fin des années 90, pour réduire l'espace de représentation.

(Deerwester et al., 1990) exposent la méthode de réduction de l'espace de représentation LSA/LSA permettant de modéliser les relations liant les mots du vocabulaire, pour en extraire des ensembles de mots appelés *classes de mots* ou *concepts*. LSA utilise une décomposition en valeurs singulières (*Singular-value decomposition* (SVD)) (Golub et Van Loan, 1989) permettant la création d'un espace sémantique représentant les relations thématiques entre les mots et les documents. Une description détaillée de ce modèle est disponible dans l'annexe B.

La méthode d'indexation sémantique latente (LSA) a été appliquée dans de nombreuses tâches en recherche d'information où elle y obtient des résultats souvent significativement meilleurs que ceux obtenus par les méthodes classiques. Dans (Dumais, 1991), des techniques similaires à celles utilisées dans les vecteurs à base de fréquence de mots ont été appliquées à la méthode LSA, comme la pondération ou l'utilisation du *relevance feedback*. Ces améliorations ont permis d'améliorer les résultats obtenus avec la méthode LSA.

Dans (Foltz et Dumais, 1992), les auteurs comparent la méthode LSA avec d'autres méthodes lors d'une tâche de filtrage d'informations. Les résultats obtenus montrent que LSA permet de sélectionner l'information la plus pertinente et de supprimer la redondance.

L'auteur dans (Dumais, 1993) a également utilisé LSA durant la campagne d'évaluation TREC-1 avec des résultats convenables. Il a montré que la taille de la représentation des documents contenus dans le corpus fourni lors de cette campagne pouvait être réduite par une décomposition SVD dans un délai raisonnable (un jour avec un seul processeur) sachant les moyens matériels limités de l'époque (1993). Dans (Dumais, 1994), l'auteur a ensuite appliqué LSA à la tâche de *question-réponse* (extraction d'information dans une base de grande taille à partir d'un jeu de requêtes) durant la campagne TREC-2 avec des résultats contrastés, principalement dus à des problèmes lors de l'utilisation du système SVD SMART (Buckley et al., 1993).

LSA a également montré son efficacité dans le domaine de la reconnaissance de la parole, et plus précisément, dans la constitution d'un modèle de langage. Ainsi, les auteurs dans (Gildea et Hofmann, 1999) proposent une méthode fondée sur LSA et l'algorithme Algorithme Espérance-Maximisation ou *Expectation-maximisation algorithm* (EM), permettant de baisser la perplexité d'un modèle n-gramme (Brown et al., 1992). La perplexité a pour vocation de mesurer la qualité de ces modèles de langage, qui sont indispensables lors de la phase de reconnaissance de la parole, en utilisant l'historique d'un terme sachant un corpus de textes. D'autres études ont confirmé l'efficacité de LSA par la suite (Berry et al., 1995; Hofmann, 1999a; Landauer et Dumais, 1997; Landauer et al., 1997; Story, 1996) dans des contextes et des tâches similaires.

Plus récemment, la méthode LSA a été utilisée dans des tâches de reconnaissance de la parole ou d'analyse des contenus parlés, par exemple pour le choix du meilleur nombre de n-grammes (n mots qui se suivent) dans les modèles de langage (Bellegarda,

2000). Le modèle LSA a été utilisé pour la détection de mots hors-vocabulaire dans des documents audios (Lecouteux et al., 2009).

Dans (Papadimitriou et al., 1998), les auteurs analysent les forces et faiblesses de LSA en développant un modèle génératif probabiliste s'appuyant sur un corpus de textes. Ils montrent que l'utilisation de LSA en lieu et place d'une méthode générative probabiliste, telle que le maximum de vraisemblance ou les méthodes bayésiennes, n'est pas forcement pertinente. Le modèle LSA ne reproduit pas convenablement la structure statistique du corpus d'apprentissage. De plus, la décomposition en valeurs singulières (SVD) nécessaire est coûteuse en ressources matérielles.

1.3.2 Analyse sémantique latente probabiliste (PLSA)

Pour répondre aux faiblesse du modèle LSA, (Hofmann, 1999a) propose une approche probabiliste de LSA appelée modèle Analyse Sémantique Latente Probabiliste ou *Probabilistic Latent Semantic Analysis* (PLSA). Ce modèle permet d'obtenir une représentation plus réaliste du document en associant plusieurs thèmes à un même document, avec des pondérations. Dans PLSA chacun des termes composant le document est associé à un thème. Le principe de ce nouveau modèle est de représenter chaque mot contenu dans la collection de documents comme un échantillon issu d'un modèle de mixtures de variables aléatoires déterminées à partir d'une loi multinomiale. Ces mots peuvent être considérés comme des "concepts" même si la relation entre cette distribution au sein du vocabulaire et un "concept", comme nous l'entendons, n'est pas explicite. Ainsi, chaque mot est généré depuis un concept ou thème, et les mots composant un document peuvent être issus de thèmes différents. Les documents sont alors représentés comme une distribution parmi les thèmes fixés composant le modèle PLSA. Cette représentation du document est dite "réduite".

Ce modèle, même s'il représente une amélioration notable de LSA, comporte néanmoins certaines faiblesses puisqu'il ne fournit pas de modèle probabiliste au niveau du document. Ainsi, PLSA considère chacun des documents composant un corpus comme un ensemble de proportions du mélange de thèmes, mais ne fournit aucun modèle génératif probabiliste de ces valeurs. Ceci entraîne deux problèmes importants :

- le nombre de paramètres grandit proportionnellement avec la taille du corpus, et donc la taille du vocabulaire, ce qui a tendance à conduire au phénomène de surapprentissage,

- la difficulté d'associer, à un document n'apparaissant pas dans le corpus à l'origine du modèle PLSA, une distribution sur les mixtures de thèmes.

Une description détaillée de la méthode PLSA/PLSA est fournie dans l'annexe C.

Bien qu'originalement pensée pour répondre aux faiblesses de la méthode LSA dans le contexte de la recherche d'information (Niu et Shi, 2010; Kim et al., 2003), cette

méthode a connu un grand succès dans d'autres domaines, tels que le traitement de l'image (Zhuang et al., 2009; Sivic et al., 2005; Fergus et al., 2005; Cao et Fei-Fei, 2007; Wong et al., 2007; Lu et Zhai, 2008; Niebles et al., 2008; Mei et al., 2007; Zhang et Gong, 2010) ou le traitement vidéo.

1.3.2.1 PLSA pour la recherche d'information

L'approche PLSA pour la représentation abstraite d'un document est encore très utilisée (Xue et al., 2008) dans diverses tâches de recherche d'information (Niu et Shi, 2010; Kim et al., 2003) et de traitement du langage naturel écrit.

(Niu et Shi, 2010) introduisent la notion de *paire de documents* dans le processus d'estimation du modèle PLSA. Ils utilisent ainsi cette relation entre paires de documents pour améliorer les processus de recherche d'information et de filtrage.

Le modèle traditionnel, à base de fréquence de mots, est comparé dans (Niu et Shi, 2010) au modèle PLSA pour les tâches de classification et de filtrage en utilisant une méthode de représentation de documents textuels à base de réseaux de neurones (*Neural Network for Text Representation* ou NNTR). Cette représentation est héritée du modèle de langage à base de réseaux de neurones probabilistiques (Bengio et al., 2006) (*Neural Probabilistic Language Model* ou NPLM).

Cette méthode a été déclinée dans la recherche d'information à plusieurs reprises pour l'adapter aux besoins propres à la tâche ou problématique étudiée. Ainsi (Hofmann, 2001) proposent une extension de PLSA qui intègre le contenu du document ainsi que les liens hypertextes dans une tâche de catégorisation de documents. Plus récemment, (Xue et al., 2008) introduisent une variante de PLSA dans une tâche de classification inter-domaine (*cross-domain*). Cette méthode, appelée *Topic-bridged* PLSA ou *Link-PLSA*, intègre dans l'apprentissage du modèle les relations thématiques entre documents étiquetés ou non, pour classifier des documents partageant des domaines communs. Ce modèle définit un processus génératif non seulement pour le contenu textuel du document, mais également pour les citations ou hyper-liens contenus dans ce document. Une suite à ces travaux est proposée dans (Nallapati et Cohen, 2008), où les auteurs proposent de fusionner le modèle *Link-PLSA* a une autre méthode substituant le modèle PLSA par le modèle d'analyse latente de Dirichlet (Blei et al., 2003; Erosheva et al., 2004) (Allocation Latente de Dirichlet ou *Latent Dirichlet Allocation* (LDA)) développé dans la section suivante. Cette dernière méthode, appelée *Link-LDA*, est donc couplée à la méthode *Link-PLSA*, maintenant nommée *Link-PLSA-LDA*, pour extraire les thèmes associés à un blog. Le modèle *Link-PLSA* a également inspiré (Chikhi et al., 2010) pour l'élaboration d'un modèle probabiliste pour l'identification automatique des structures entre communautés (ISC). Ce système, dénommé *Smoothed Probabilistic Community Explorer* ou SPCE, met en œuvre une technique de lissage pour la tâche de ISC.

1.3.2.2 PLSA pour le traitement de la parole

Un domaine, proche du traitement automatique du langage écrit, est le traitement du langage parlé. Là aussi, PLSA a démontré sa capacité à fournir des caractéristiques (ou *features*) permettant d'augmenter la robustesse des systèmes de reconnaissance automatique de la parole (SRAP) (voir annexe E pour plus de détails sur ces systèmes). Par exemple (Mrva et Woodland, 2004) utilisent PLSA pour diminuer la perplexité d'un modèle de langage pour la Reconnaissance Automatique de la Parole (RAP). Ce modèle est nécessaire dans la phase de reconnaissance d'un terme sachant son historique dans un corpus d'entraînement. Les auteurs introduisent, dans le calcul de la probabilité d'un terme au sein d'un n-gramme, la probabilité obtenue lors de l'apprentissage du modèle PLSA via l'algorithme d'*espérance-maximisation* (EM) (voir annexe C pour de plus amples informations concernant cet algorithme et le mode de calcul des probabilités de PLSA).

L'utilisation de PLSA pour améliorer les modèles de langage est également proposée par (Akita et Kawahara, 2004) dans un contexte de réunion (*meeting*) où plusieurs personnes discutent de thèmes divers. Les auteurs adaptent le modèle de langage traditionnel à base de n-grammes, en y introduisant les caractéristiques de l'intervenant et les thèmes de discussion. Le modèle PLSA est combiné à un modèle d'unigrammes (Gildea et Hofmann, 1999) qui évalue la pertinence d'un tel système en termes de perplexité du modèle de langage et de pertinence des mots contenus dans la discussion. Les résultats montrent que le système proposé, utilisant un modèle thématique issu de PLSA, obtient de meilleurs résultats en termes de perplexité et de mots reconnus dans le cadre de réunions multi-thèmes.

1.3.2.3 PLSA pour le traitement de l'image

De manière identique aux tâches de recherche d'information telles que la classification de documents textuels, le paradigme PLSA est utilisé avec succès dans la tâche de classification d'images. Dans (Zhuang et al., 2009), les auteurs utilisent une partie de l'information des étiquettes des documents associées aux images durant la phase d'apprentissage du modèle, et introduisent la notion de seuillage de la probabilité entre thèmes et documents durant cette même phase.

(Sivic et al., 2005) emploient le modèle PLSA pour trouver aussi bien la catégorie d'un objet que sa grille spatiale de manière non-supervisée. Dans le modèle proposé, chacune des catégories correspond à un thème donné et chaque mot est considéré comme un point d'intérêt. D'autres méthodes issues du domaine du traitement de l'image ont essayé d'inclure l'aspect spatial d'une partie de l'image dans le modèle PLSA comme (Fergus et al., 2005). Ces chercheurs ont incorporé une information spatiale des pièces d'une image dans le modèle PLSA pour améliorer la recherche de catégories d'une image donnée dans les moteurs de recherche. Dans (Cao et Fei-Fei, 2007), les auteurs proposent un espace de thèmes spatial cohérent (*Spatial-LTM*) pour la segmentation et la reconnaissance d'objets ainsi que des scènes. Ce modèle est appris

dans un contexte aussi bien supervisé que non-supervisé.

Dans (Niebles et al., 2008), les zones locales sont extraites par un système de détection de zones d'intérêt en trois dimensions, comme des mots dans les modèles PLSA. Ainsi, ces deux modèles permettent d'assigner, à une zone d'intérêt, un thème correspondant à une action d'une manière totalement non-supervisée. Cette méthode permet de ranger une vidéo dans une catégorie d'action et, ainsi, de localiser l'action dans une vidéo. Le modèle proposé par (Niebles et al., 2008) ignore la disposition spatiale relative de zones d'une image. Sachant que les déplacements humains sont de caractère temporel et dynamique, cette disposition est une donnée importante dans la tâche de détection de mouvements chez l'homme. Cette faiblesse est traitée dans (Wong et al., 2007) où les auteurs étendent PLSA pour capturer indistinctement la structure et l'information sémantique d'une zone d'intérêt locale pour reconnaître et localiser une action humaine. De plus, l'approche PLSA ne nécessite pas une décomposition en valeurs singulières, qui est souvent coûteuse en termes de temps de calcul.

D'autres études incorporent l'étiquette de la classe durant l'apprentissage du modèle PLSA. C'est le cas de (Lu et Zhai, 2008), où les auteurs définissent une distribution conjuguée *a priori* (Mei et al., 2007) pour des segments d'experts, et proposent un modèle semi-supervisé PLSA pour l'analyse de blogs.

Plus récemment, (Zhang et Gong, 2010) ont traité les *frames* comme des mots et ont développé un modèle PLSA structuré pour prendre en compte les dépendances temporelles des mots pour la catégorisation d'actions humaines. Ils démontrent dans leur étude que la méthode PLSA est un cas particulier de sa version structurée. Cette version structurée de PLSA est apprise de manière non-supervisée.

Les travaux de Hofmann ont permis une avancé majeure dans le domaine de la représentation de documents dans un espace thématique en permettant de formaliser les relations statistiques entre documents en tenant compte des caractéristiques des données, telles que la fréquence des mots (ou TF-IDF).

Les modèles issus de PLSA sont néanmoins incomplets dans la mesure où ils ne fournissent pas de probabilités au niveau du document. Dans la méthode PLSA, chacun des documents est défini par un ensemble de valeurs représentant les mélanges des proportions des thèmes. Il n'y a aucun modèle génératif probabiliste pour générer ces valeurs composant l'image du document dans l'espace thématique, ce qui pose deux problèmes majeurs :

- Le nombre de paramètres grandit avec la taille du corpus et conduit au phénomène bien connu du surapprentissage.
- L'attribution d'une probabilité d'un thème sachant un document n'est pas claire.

Le modèle fondé sur l'analyse latente de Dirichlet (*Latent Dirichlet Allocation* (LDA)) répond à ces deux verrous que sont le surapprentissage et l'estimation de la probabilité d'un document étranger au corpus d'apprentissage du modèle PLSA. Il propose un environnement plus adapté aux tâches de recherche d'information, bien qu'il puisse

être considéré comme voisin de PLSA (Girolami et Kabán, 2003).

1.3.3 Allocation latente de Dirichlet (LDA)

Le modèle LDA, proposé pour la première fois par (Blei et al., 2003), offre une solution pour contourner les défauts du modèle PLSA. Le nombre de paramètres du modèle est limité par le nombre k de thèmes le composant. Ainsi, le modèle conserve une taille fixe n'augmentant pas avec le nombre de documents composant le corpus d'apprentissage. Le paradigme LDA permet également une bonne estimation de la probabilité d'un document non rencontré lors de la phase d'entraînement, connaissant les thèmes composant le modèle. Il est explicité plus précisément dans l'annexe D.

Le modèle LDA est minutieusement décrit dans (Blei et al., 2003) ou dans (Griffiths et Steyvers, 2004; Heinrich, 2005; Blei et Lafferty, 2009; Berry et Kogan, 2010).

L'impact de la méthode LDA dans le domaine du traitement automatique de la langue écrite (puis dans d'autres domaines par la suite comme l'image et l'audio) est majeur (Wang et McCallum, 2005).

LDA a suscité de nombreux travaux qui s'intéressent aux problèmes d'estimation sous-jacents (Griffiths et Steyvers, 2004; Asuncion et al., 2009) ou proposent des extensions de ce modèle :

- le processus hiérarchique de Dirichlet ou *hierarchical Dirichlet processes* (HDP) (Teh et al., 2004),
- le modèle dynamique de thèmes ou *dynamical topic model* (DTM) (Blei et Lafferty, 2006b),
- le modèle corrélé de thèmes ou *correlated topic model* (CTM) (Blei et al., 2007),
- le modèle auteur-thème LDA ou *Author-Topic LDA* (Rosen-Zvi et al., 2004),
- ou *labeled LDA* : un modèle plus récent utilisant les étiquettes attribuées à un document composant le corpus d'apprentissage du modèle (Ramage et al., 2009). Ce modèle propose d'utiliser la connaissance *a priori* de classes auxquelles les documents appartiennent pour élaborer des espaces de thèmes "guidés" par ces étiquettes.

Ces variantes ne sont pas détaillées ici, cette partie se concentrant sur le modèle initial (LDA) utilisé dans la suite de ce travail.

L'estimation des paramètres (distributions) du modèle LDA n'est pas triviale comme nous pouvons le constater dans l'équation D.9. En effet, il est nécessaire de réaliser la somme sur toutes les combinaisons possibles d'attribution des thèmes z

($n_{d,k,.}, n_{.,k,v}$). Ceci rend le calcul des probabilités impossible (Blei et al., 2003). L'utilisation de la méthode classique pour l'estimation des paramètres d'un modèle telle que l'algorithme EM (voir section C.3) est à éviter. Cependant, les méthodes provenant du *machine learning* permettent de contourner cette difficulté. Ces méthodes sont décrites dans la section suivante.

1.3.3.1 Algorithmes d'estimation des paramètres du modèle LDA

Plusieurs algorithmes (Asuncion et al., 2009; Blei et Lafferty, 2009), issus du domaine du *machine learning*, ont été utilisés pour estimer les paramètres du modèle LDA :

- **Algorithme de maximum de vraisemblance**, ou *Maximum likelihood* (ML), est une méthode d'estimation des paramètres de l'algorithme PLSA (Hofmann, 1999a) où les paramètres α et β sont ignorés alors que les paramètres θ et ϕ sont traités séparément. PLSA est un modèle génératif probabiliste, semblable à LDA, dépourvu de la distribution de Dirichlet. L'algorithme **de Maximum a posteriori** (Chien et Wu, 2008) est une approche semblable au modèle PLSA de Hofmann.

- **Algorithme d'inférence bayésienne variationnelle ou** *Variational Bayesian Inference* **(VBI)**, ou (*Variational Bayesian* (VB)) (Blei et al., 2003) dans la version lissée du modèle LDA, est une méthode dont les hyper-paramètres z, θ et ϕ sont des variables latentes où les distributions *a posteriori* sont estimées en utilisant la distribution variationnelle bayésienne. Cette méthode a été étendue par (Griffiths et Steyvers, 2004) dans une version dite *collapsed*, c'est-à-dire que l'estimation d'un paramètre A dans un ensemble de paramètres A, B et C ne se fera plus dans sa version conditionnelle (échantillonnage pour $P(A|B,C)$ mais dans sa version marginalisée). Celle-ci sera alors échantillonnée selon la marginalisation de $P(A|B,C)$:
$$P(A|B,C) = \int_B P(A|C)dB$$

- **Collapsed Gibbs Sampling** (CGS) est un algorithme bâti sur les *Markov Chain Monte-Carlo* (MCMC) explicité plus en détails dans la section suivante. Il se caractérise par une convergence rapide vers une configuration optimale ($\sim$ 100 itérations).

1.3.3.2 *Collapsed Gibbs Sampling*

Les variables latentes du modèle LDA sont difficiles à estimer directement (voir équation D.9 de l'annexe D) et cette complexité est due à la taille du corpus (w) combinée au nombre de combinaisons à évaluer. Le *Collapsed Gibbs Sampling* (Griffiths et Steyvers, 2004) est une méthode utilisant un algorithme d'échantillonnage permettant d'es-

timer les paramètres d'un espace discret de grande dimension (Steyvers et al., 2004). Les auteurs ont appliqué l'algorithme du *Gibbs Sampling* pour la première fois dans le cadre de l'approximation des paramètres du modèle LDA. Ce procédé d'approximation est inspiré du MCMC. Cette méthode résout les problèmes d'échantillonnage depuis des distributions complexes de probabilités en utilisant des variables aléatoires (MacKay, 2003). (Carpenter, 2010) définit le contexte d'application du Gibbs sampling pour l'estimation des paramètres du modèle LDA, comme la nécessité d'estimer la probabilité qu'un thème $z_{a,b}$ soit assigné au terme $w_{a,b}$ ($a^{\text{ème}}$ terme du vocabulaire du $b^{\text{ème}}$ document), sachant tous les thèmes assignés ç toutes les autres occurrences du corpus $z_{-(a,b)}$:

$$\begin{aligned} P(z_{a,b}|z_{-(a,b)}, w_{a,b}, \alpha, \beta) &\propto \int_{\theta}\int_{\phi} P(w, z, \theta, \phi|\alpha, \beta)\mathrm{d}\theta\mathrm{d}\phi \\ &\propto \frac{n^{-(a,b)}_{z_{a,b},a,.} + \alpha}{n^{-(a,.)}_{z_{a,b},a,.} + K\alpha} \times \frac{n^{-(a,b)}_{z_{a,b},.,w_{a,b}} + \beta}{n^{-(a,b)}_{z_{a,b},.,.} + |V|\beta}, \end{aligned} \tag{1.1}$$

En réalisant l'hypothèse que les variables α et β sont asymétriques et que le dénominateur lié au thème est omis car il ne dépend pas des mots du vocabulaire, l'équation 1.1 devient :

$$P(z_{a,b}|z_{-(a,b)}, w, \alpha, \beta) \propto \frac{\left(n^{-(a,b)}_{z_{a,b},a,.} + \alpha_{z_{a,b}}\right) \times \left(n^{-(a,b)}_{z_{a,b},.,w_{a,b}} + \beta_{w_{a,b}}\right)}{n^{-(a,b)}_{z_{a,b},.,.} + |V| \times \beta_j}, \tag{1.2}$$

avec $n^{-(a,b)}_{z_{a,b},a,.}$ le nombre de termes (hormis w) contenus dans le document b qui sont attribués au thème $z_{a,b}$; le terme $n^{-(a,b)}_{z_{a,b},.,w_{a,b}}$ correspond au nombre de fois où le terme $w_{a,b}$ est assigné au thème $z_{a,b}$. Le dénominateur $n^{-(a,b)}_{z_{a,b},.,.} + \sum_{j=1}^{J} \beta_j$ a pour unique vocation de normaliser le second terme du numérateur :

$$P(z_{a,b}|z_{-(a,b)}, w_{a,b}, \alpha, \beta) = \frac{\left(\frac{\left(n^{-(a,b)}_{z_{a,b},a,.}+\alpha_{z_{a,b}}\right)\times\left(n^{-(a,b)}_{z_{a,b},.,w_{a,b}}+\beta_{w_{a,b}}\right)}{n^{-(a,b)}_{z_{a,b},.,.}+|V|\times\beta_j}\right)}{\left(\sum_{k=1}^{K}\frac{\left(n^{-(a,b)}_{k,a,.}+\alpha_k\right)\times\left(n^{-(a,b)}_{k,.,w_{a,b}}+\beta_{w_{a,b}}\right)}{n^{-(a,b)}_{k,.,.}+|V|\times\beta_j}\right)}. \tag{1.3}$$

Les variables θ et ϕ du modèle LDA sont estimées comme suit (Griffiths et Steyvers, 2004) :

$$\hat{\theta}_{b,k} = \frac{\alpha_k + n_{b,k,.}}{\alpha_. + n_{b,.,.}} \tag{1.4}$$

$$\hat{\phi}_{k,v} = \frac{\beta_{k,v} + n_{.,k,v}}{\beta_{k,.} + n_{.,k,.}}. \tag{1.5}$$

L'algorithme permettant la production des échantillons de la distribution à postériori des variables latentes (z) à partir desquelles on pourra estimer θ pour tous les documents et ϕ pour tous les thèmes est donné dans 1.

Algorithm 1: Échantillonage pour l'estimation des matrice θ ainsi que ϕ du modèle LDA avec la méthode du *Gibbs Sampling*.

Data: Corpus de J documents composé d'un vocabulaire de taille N
Result: $\hat{\theta}_{d,k}$ et $\hat{\phi}_{k,v}$
for $b \leftarrow 1$ **until** J **do**
  for $a \leftarrow 1$ **until** N **do**
    u $\leftarrow$ tirer une valeur selon une loi [0,1]
    for $k \leftarrow 1$ **until** K **do**
      $$P(k) \leftarrow P(k-1) + \frac{\left(n_{k,a,.}^{-(a,b)} + \alpha_k\right) \times \left(n_{k,.,w_{a,b}}^{-(a,b)} + \beta_{w_{a,b}}\right)}{n_{k,.,.}^{-(a,b)} + |V| \times \beta_j}$$
    end
    for $k \leftarrow 1$ **until** K **do**
      if $u < \frac{P(k)}{P(K)}$ **then**
        $z_{a,b} = k$, stop
      end
    end
  end
end

La section suivante présente les principales méthodes d'évaluation des modèles thématiques, et plus particulièrement le modèle LDA (Wallach et al., 2009).

1.3.3.3 Mesure de la performance des modèles à base de thèmes

Un modèle bâti sur des espaces de thèmes comme LSA, PLSA ou LDA entre autres, nécessite un corpus de documents hétérogènes de grandes tailles caractérisant de manière diversifiée les contextes possibles où un mot, ou une succession de mots, peuvent être rencontrés. L'objectif premier de tels modèles est d'inférer des documents n'apparaissant pas dans les données d'apprentissage. Évaluer la généricité du modèle est un problème critique et de nombreux travaux se sont intéressés à la validation empirique de ce type de modèle. (Rosen-Zvi et al., 2004; Wallach et al., 2009) proposent de découper les documents du corpus en deux, constituant ainsi le corpus d'apprentissage et le corpus de validation ou de test pour élaborer ces deux sous-ensembles. Cette méthode ne permet néanmoins pas de conserver la qualité individuelle de chacun des documents (Buntine, 2009). Les chercheurs du domaine du traitement automatique de la langue et de la RI divisent donc le plus souvent les données disponibles pour l'apprentissage d'un modèle en deux sous-ensembles dénommés données d'entraînement et données de validation (ou de test). (Griffiths et Steyvers, 2004) utilisent, eux, l'ensemble du corpus pour évaluer la pertinence du modèle proposé.

Ces métriques d'évaluation des modèles permettent, par exemple dans le cas du modèle LDA, de convenir d'un nombre pertinent de thèmes composant l'espace de représentation en minimisant la perplexité dudit modèle (Blei et Lafferty, 2009).

La perplexité est la métrique d'évaluation de modèles la plus utilisée. Celle-ci décroît lorsque le log-vraisemblance du modèle augmente. Une perplexité faible indique un modèle au pouvoir de généralisation plus élevé (Blei et al., 2003; Rosen-Zvi et al., 2004). La perplexité équivaut à la moyenne géométrique inverse de la vraisemblance par terme composant le corpus d'évaluation :

$$\text{perplexité}(\mathcal{B}) = exp\left\{-\frac{1}{N_{\mathcal{B}}}\sum_{d=1}^{M}\log P(\mathbf{w})\right\} \tag{1.6}$$

avec

$$N_{\mathcal{B}} = \sum_{d=1}^{M} N_d\ , \tag{1.7}$$

où $N_{\mathcal{B}}$ est la longueur des N documents ; M l'ensemble des termes ; N_d est le nombre de mots contenu dans le document d ; $P(\mathbf{w})$ est la vraisemblance que le modèle génératif assigne, à un document d du corpus d'évaluation, un terme $\mathbf{w}$. La quantité contenue dans l'exponentielle est appelée entropie des données d'évaluation sachant le modèle. L'utilisation du logarithme permet d'interpréter cette entropie en termes de *bits* d'information.

La perplexité est la métrique d'évaluation standard dans le domaine du traitement de la langue. D'autres métriques existent, comme *Empirical likelihood* (Li et McCallum, 2006), dérivée de la perplexité. Une autre famille de métriques est la vraisemblance marginale (*marginal likelihood*), qui a donné lieu à de nombreux de travaux et donc, de nouvelles métriques (Chib, 1995; Murray et Salakhutdinov, 2008; Del Moral et al., 2006; Buntine, 2009; Newton et Raftery, 1994; Griffiths et Steyvers, 2004; Griffiths et al., 2004).

1.3.3.4 LDA dans le domaine du traitement automatique du langage écrit

Historiquement, le modèle génératif LDA a été employé dans le domaine de la recherche d'information (RI), puis dans le traitement automatique du langage écrit ou pour le traitement de la parole. Le domaine de la RI regorge de champs d'application possibles dans lesquels le modèle LDA trouve un intérêt certain, compte tenu :

- du faible nombre de paramètres à estimer qu'il nécessite,
- de la facilité apparented'apprentissage,
- de son pouvoir d'inférer de nouveaux documents.

Ainsi, un problème central en RI, qui a bénéficié du modèle LDA et de ses diverses variantes, est la recherche de documents pertinents sachant une requête donnée. Dans (Wei et Croft, 2006), les auteurs proposent une première étude du modèle LDA lors d'une tâche de recherche de documents pertinents au sein d'un large corpus, répondant à une requête dans le cadre de la campagne Text REtrieval Conference (TREC). Ils étendent le modèle de maximum de vraisemblance, permettant de déterminer la probabilité *a posteriori* d'un terme sachant une collection de documents (Zhai et Lafferty, 2001), en y incorporant la probabilité d'un terme sachant une collection issue du modèle LDA. Cette probabilité est déterminée en fonction de la probabilité du terme au sein des thèmes composant l'espace de thèmes LDA et la probabilité de chacun des thèmes sachant le document. Dans (Azzopardi et al., 2004), les auteurs utilisent le modèle LDA sur des corpus de taille réduite (100 documents comparativement à ceux utilisés aujourd'hui comme Wikipedia ou TREC) pour l'extraction de termes au sein d'un corpus hétérogène (issu de trois domaines : médecine, aéronautique et science de l'information).

Une variante du *Batch Variational Bayesian* (BVB) (VB-LDA) (Blei et al., 2003) pour l'apprentissage d'un modèle LDA est proposée par (Hoffman et al., 2010) avec une version *online* (*online* VB-LDA). Cette version permet de diminuer la perplexité du modèle comparativement à un modèle LDA "classique". VB-LDA permet également d'ajouter "en ligne" des documents dans le modèle lors de la phase d'apprentissage. Ainsi, cette méthode peut s'appliquer aisément en modifiant simplement l'algorithme d'inférence des paramètres du modèle. La méthode VB-LDA est également applicable à d'autres modèles hiérarchiques bayésiens.

1.3.3.5 LDA dans le domaine de la parole

La méthode LDA a été utilisée dans les tâches de classification ou de catégorisation de documents textuels principalement. Dans le domaine du traitement automatique de la parole, l'approche LDA a été implémentée pour l'adaptation du modèle de langage (Heidel et al., 2007). Les auteurs l'utilisent pour déterminer la probabilité qu'un document soit généré sans tenir compte de l'ordre des mots. De plus, LDA construit son modèle de thèmes au niveau du document, ce qui est contraire à l'idée initiale des modèles de langage dans le domaine de la reconnaissance automatique de la parole, qui eux sont fondés sur le modèle n-gramme. L'auteur dans (Wallach, 2006) propose de ne pas utiliser le document comme un ensemble de termes isolés, mais comme un ensemble de bi-grammes. Ainsi, le modèle de langage initial bâti sur le mot et son "contexte" (n-gramme) est préservé. Le document est ainsi vu comme un "sac-de-bi-grammes".

Comme évoqué précédemment, LDA est une méthode voisine de PLSA. Ces deux techniques sont comparées pour l'élaboration d'un modèle de langage performant en termes de Taux d'erreur-mot (TEM) ou *Word Error Rate* (WER) dans (Chien et Chueh, 2008). Dans cet article, les auteurs proposent un Modèle de Langage Latent de Dirichlet (MLLD) ou *latent Dirichlet language model* (LDLM) pour la modélisation de séquences de mots n-grammes du modèle de langage. Le modèle LDLM est alors comparé à un

modèle de langage utilisant PLSA dans la tâche de reconnaissance de la parole sur le corpus du *The Wall Street Journal* (WSJ). Les résultats montrent que le modèle issu d'une analyse latente de Dirichlet obtient de meilleurs résultats (TEM de 5,19 %) que le système de base (modèle n-gramme) avec un TEM de 5,38 %, ou que le modèle de langage proposé par (Gildea et Hofmann, 1999) s'appuyant sur l'algorithme EM (Annexe C.3) avec un TEM de 5,25 %. Plusieurs autres méthodes utilisant LDA pour composer un modèle de langage ont été proposées comme le *Topic Cache Latent Dirichlet language model* (TCLDLM) (Chueh et Chien, 2010) ou le *Dirichlet class language model* (DCLM) (Chien et Chueh, 2011).

D'autres études sont allées plus loin en proposant ce dernier modèle de langage (DCLM) fondé sur le modèle LDA, mais en tenant compte de l'historique d'un terme. Cette séquence de termes est projetée dans un espace de thèmes pour déterminer le log-vraisemblance marginal sur les classes apprises par le modèle LDA. Ce modèle ne considère plus le document comme un "sac-de-mots", mais comme un ensemble de termes combinés à leurs historiques dans le modèle de langage s'appuyant sur les thèmes latents. Ce procédé permet de déterminer les séquences de n-grammes d'une manière automatique non-supervisée. Cette méthode a montré son efficacité dans la tâche de reconnaissance de la parole sur le même corpus (WSJ) comparativement à d'autres méthodes fondées sur les espaces de thèmes comme LDLM avec un TEM de 5,02 %, atteignant même un TEM de 4,92 % pour la version dite *cache* de DCLM.

Toutes ces études sont motivées par la faiblesse de l'information sémantique intégrée au Système de Reconnaissance Automatique de la Parole (SRAP). Malgré des progrès constatés dans des situations particulières, LDA n'est pas devenu une méthode "consensuelle" en RAP mais elle est devenue un outil standard de l'analyse des contenus parlés (*Speech Analytics*).

1.3.3.6 LDA dans le domaine du traitement de l'image

Dans le domaine du traitement de l'image, l'approche LDA a été utilisée avec le même succès que dans les domaines du traitement automatique du langage, qu'il soient écrit ou oral. Ainsi, les auteurs dans (Iwata et al., 2007) utilisent une méthode de visualisation de probabilités apprises *a posteriori* pour décrire les relations entre les classes (ou concepts) issues de LDA dans une représentation graphique claire. Les concepts issus de LDA ont trouvé un écho dans la tâche d'annotation d'images satellites (Lienou et al., 2010). Dans cette étude, les auteurs combinent l'information extraite des images satellites et les concepts issus de LDA, pour annoter ces images. Ainsi, les auteurs montrent que l'utilisation de caractéristiques issues de LDA, telles que la moyenne et l'écart-type de chacun des mots (fenêtre de 10 par 10 dans l'image satellite) composant l'ensemble de données, permet d'améliorer les résultats.

Le modèle LDA est également utilisé dans la tâche de catégorisation d'images (Russell et al., 2006; Cao et Fei-Fei, 2007; Fei-Fei et Perona, 2005). Par exemple, les auteurs dans (Fei-Fei et Perona, 2005) utilisent LDA pour retrouver des scènes et de les grouper. Cette méthode ne nécessite pas de processus d'étiquetage d'un ensemble d'images

d'apprentissage par un expert. Un ensemble de caractéristiques locales des images du corpus d'apprentissage est, dans un premier temps, extrait. Ces caractéristiques sont nécessaires pour caractériser les éléments contenus dans ces images. Dans une seconde étape, ces caractéristiques sont utilisées comme un sac-de-mots depuis lesquelles, un ensemble de classes (ou *clusters*) est appris en utilisant la technique LDA. Ces regroupements (*clusters*) de formes seront employés pour catégoriser ou regrouper des images en fonction des formes qui la composent.

Les applications de modélisation d'activités et d'interactions dans une scène fortement peuplée ont également montré la pertinence de l'utilisation d'un modèle issu d'une allocation latente de Dirichlet. Dans (Wang et al., 2007), les auteurs utilisent ce modèle pour regrouper les "mots visuels" de bas niveau, comme les zones d'intérêt ou des pixels en mouvement, dans des espaces de thèmes correspondant à des objets, actions humaines ou des faibles activités en utilisant leurs co-occurrences dans les images.

1.4 Conclusions

Le modèle LDA permet d'obtenir une représentation des documents dans un espace latent de taille réduit. Cette représentation basée sur un espace de thèmes LDA à l'avantage de permettre d'inférer des documents n'apparaissant pas dans le corpus d'apprentissage dans cet espace et le nombre de paramètres du modèle est réduit. La faiblesse majeur d'une telle représentation, est le choix des hyper-paramètres (α, β et le nombre de classes composant le modèle) qui est souvent ardu.

Chapitre 2

Représentation robuste de documents par projection dans un espace thématique homogène

Sommaire

Résumé

Dans ce chapitre, nous évaluons l'intérêt ainsi que le potentiel de la représentation "sémantique" obtenue par analyse latente de Dirichlet (LDA) pour le traitement de documents bruités, d'abord sur une tâche de catégorisation puis sur un ensemble de problèmes de RI.

2.1 Introduction

Les méthodes de représentation d'un document présentées dans le chapitre précédent ont montré leur efficacité dans des domaines variés tels que le traitement d'images, de documents audios ou textuels. Ceci est particulièrement le cas pour l'approche LDA qui a été largement utilisée pour la représentation thématique d'un document. Ce succès est dû à sa capacité à extraire la structure sémantique sous-jacente d'une collection de documents, à permettre des représentations de haut-niveau ne nécessitant qu'un nombre réduit de paramètres et, de façon générale, à son pouvoir de généralisation important.

Ce chapitre a pour objectif d'évaluer ce que l'allocation latente de Dirichlet (LDA) peut apporter en termes de robustesse à divers types de bruit. Dans la section 2.2, nous évaluons les performances de ce type de représentation sur des documents parlés souvent mal transcrits (robustesse aux erreurs de transcriptions). Dans la section 2.3, nous évaluons la robustesse d'une représentation s'appuyant sur un espace de thèmes LDA à des styles d'écritures atypiques issus du Web.

2.2 Représentation vectorielle robuste de documents parlés fortement bruités

La catégorisation de documents est un problème très central en recherche d'information, en fouille de texte ainsi qu'en TALN (voir section). Cette section évalue l'apport de la représentation dans des espaces thématiques de documents audios fortement bruités lors d'une tâche de catégorisation.

2.2.1 Problématiques liées à la catégorisation de transcriptions automatiques fortement bruitées

Un domaine partageant des racines communes au domaine de la recherche d'information est le traitement automatique de la parole. La reconnaissance automatique de la parole (RAP) est un domaine incontournable du traitement automatique de la parole ayant pour objectif de transcrire des documents audios issus d'enregistrements ou de vidéos. Ces transcriptions peuvent ensuite être traitées comme des documents textuels. Néanmoins, la parole est un mode d'expression particulier et les structures grammaticales impliquées ne sont pas celles des documents écrits. De plus, les SRAP (voir annexe E) peuvent commettre des erreurs liées aux mots hors-vocabulaire, aux conditions acoustiques difficiles (environnement, qualité du matériel, ...), aux écarts entre conditions d'entraînement et d'utilisation du système, aux disfluences, ... Les performances des systèmes sont variables et le fait que des séquences de termes soient mal transcrites peut nuire à la bonne compréhension du document et à des difficultés d'analyse par des méthodes s'appuyant sur les transcriptions automatiques imparfaites (catégorisation, traduction, ...).

Cette section propose de confronter deux représentations d'un document transcrit. Ces représentations sont évaluées sur une tâche de catégorisation de documents.

Une représentation de la transcription dans un espace continu de caractéristiques est généralement nécessaire pour une tâche de catégorisation (voir annexe A). Dans les tâches de recherche d'information (RI), une caractéristique fréquemment utilisée est la *fréquence de mots* pour l'extraction d'un sous-ensemble de termes discriminants [1] pour une classe ou catégorie de documents donnée. Cet ensemble de termes discriminants devrait permettre de composer un vecteur de représentation d'une transcription dans l'espace sémantique.

Bien que la caractéristique fondée sur la fréquence de termes ait démontré son efficacité dans le domaine de la RI, son application à des transcriptions automatiques de documents audios est plus complexe. En effet, une représentation du document transcrit automatiquement utilisant les termes qui la composent, peut s'avérer erronée du fait des erreurs du SRAP.

L'évaluation présentée dans cette section repose sur l'idée qu'un document audio transcrit de manière automatique peut être considéré dans un espace de représentation intermédiaire composé de thèmes, afin de pallier la difficulté de transcription que peut rencontrer le SRAP dans des conditions non optimales. Nous espérons de cette projection d'une transcription dans un espace de thèmes qu'elle augmente la robustesse des systèmes de catégorisation aux erreurs de reconnaissance.

Dans la suite, nous évaluons la représentation s'appuyant sur la fréquence de mots (TF-IDF combinée au critère de pureté de Gini (Dong et al., 2011)), avec la représentation à base d'espace de thèmes issue de LDA (Blei et al., 2003). Lors de la projection d'une transcription dans l'espace de thèmes, une réduction du vecteur de représentation du document est réalisée, en passant d'un espace contenant l'ensemble du vocabulaire à une représentation dont la taille est fixée par le nombre de classes contenu dans l'espace thématique. Cette dimension sera réduite artificiellement en utilisant une Analyse en Composantes Principales (ACP) ou *Principal Component Analysis*.

Les représentations à base de fréquence de mots ou d'un espace thématique, sont ensuite couplées à un système de catégorisation pour identifier automatiquement la catégorie associée à chacune des transcriptions automatiques fortement bruitées. Le choix d'une méthode efficace de catégorisation adaptée aux données est essentiel. Les documents composant le corpus d'entraînement, de développement, et de test, sont élaborés depuis un espace de thèmes qui ne peut être modifié. Pour cette raison, la seconde partie de cette étude se penche sur le choix d'une méthode de classification permettant de garder l'uniformité des représentations thématiques des transcriptions. Une méthode classique transformant les données à l'aide d'une fonction noyau et performante, appelée Machines à vecteurs de support ou *ou Support Vector Machine* (SVM) (Cortes et Vapnik, 1995), est comparée en termes de bonne catégorisation, à une méthode bayésienne naïve préservant l'homogénéité des données. Cette étude permet

1. Le terme "discriminant" s'applique à un terme, si celui-ci permet de distinguer une classe (catégorie) de toutes les autres.

d'évaluer la validité de deux hypothèses émises par la méthode bayésienne naïve :

- la distribution des thèmes suit une loi gaussienne,
- l'égalité entre les matrices de covariance de chacune des classes.

Un document audio transcrit automatiquement par un SRAP contient un nombre de termes mal transcrits, absents ou remplacés par d'autres termes. Pour quantifier ces erreurs, une métrique bien connue dans le domaine de la reconnaissance de la parole, est le TEM. Cette mesure, qui compare une transcription automatique avec le même document retranscrit par un humain, permet donc de mesurer la qualité d'un SRAP. Elle peut être adaptée à l'utilisation faite des transcriptions.

La qualité de la représentation d'une transcription automatique d'un document audio sera mesurée sur une tâche de catégorisation de documents transcrits. Une analyse sera proposée en fin de cette section, afin de montrer que l'impact d'une réduction de l'espace de représentation par une ACP est nulle. Nous constaterons donc que le taux de transcriptions bien catégorisées est fortement lié à l'uniformité de la représentation de la transcription dans l'espace de thèmes. Finalement, nous étudierons l'apport d'une adaptation du TEM à la tâche de catégorisation.

La section 2.2.2 présentera un survol des méthodes de représentation de contenus parlés. La représentation des transcriptions dans les deux espaces à base de fréquence de mots et d'espace de thèmes sont décrites dans la section 2.2.3. Deux méthodes de catégorisation sont comparées dans la section 2.2.4. Les sections 2.2.5 et 2.2.6 décrivent le protocole expérimental ainsi que les résultats obtenus avant de conclure dans la section 2.2.7.

2.2.2 Historique des méthodes pour la catégorisation de documents audios transcrits automatiquement

Des études récentes pour l'analyse de conversations parlées, pour l'identification du thème dominant ou encore pour la segmentation, sont disponibles respectivement dans (Tur et De Mori, 2011), (Melamed et Gilbert, 2011), (Hazen, 2011) and (Purver, 2011).

Les méthodes de représentation d'un document parlé retranscrit automatiquement par un SRAP, partagent des similitudes avec celles d'un document écrit. Toutefois, ces documents diffèrent dans leur contenu en termes de structure grammaticale et de vocabulaire.

Une méthode de représentation classique de documents écrits dans un espace vectoriel utilisant la fréquence de mots, appelée TF-IDF (Robertson, 2004), est fréquemment utilisée dans le domaine de la recherche d'information pour l'extraction de termes par exemple. Ce procédé est parfois associé au critère de pureté de Gini (Demiriz et al., 1999), par exemple pour l'extraction de mots-clés. La représentation TF-IDF est détaillée

dans l'annexe A. Ce type de méthodes permet de produire des représentations numériques de documents dans des espaces vectoriels classiques et de permettre leur traitement par les méthodes d'analyse, par exemple pour la catégorisation de documents.

La catégorisation de documents est un cas particulier de catégorisation de formes et les classifieurs employés sont, le plus souvent, issus du domaine de la classification automatique. Plusieurs approches pour la catégorisation de documents ont été étudiées dans le passé. Une des plus utilisées est la méthode de SVM (Yuan et al., 2012). Les méthodes composant les SVM sont un ensemble de techniques d'apprentissage supervisées. Connaissant un échantillon de données, la méthode à base de SVM détermine un plan séparateur entre les données appelé "vecteur support". Ensuite, un hyperplan séparateur maximisant la "marge" entre les vecteurs de support est calculé (Vapnik, 1963). Cette technique a été utilisée pour la première fois par (Boser et al., 1992) dans des tâches de régression (Müller et al., 1997) et de classification (Joachims, 1999). Le succès des techniques SVMs est dû à ses performances dans ces deux tâches, ainsi qu'au faible nombre de paramètres nécessitant un ajustement lors de la phase d'apprentissage (Bartlett et Shawe-Taylor, 1999).

La combinaison d'un espace thématique issu de LDA avec un système de catégorisation utilisant des SVMs, a été récemment utilisée dans des domaines variés comme la biologie (hua Yeh et hsing Chen, 2010), la catégorisation de textes (Zrigui et al., 2012), la recherche d'information dans des documents audios (Kim et al., 2009), la détection d'événements dans des documents multimedias (Morchid et al., 2013a) ou encore la détection de scènes dans un corpus d'images (Tang et al., 2009). Cette combinaison alliant le modèle probabiliste génératif LDA et la méthode de catégorisation à base de SVMs a également été explorée dans le contexte d'extraction de mots-clés dans des transcriptions automatiques (Sheeba et Vivekanandan, 2012). À notre connaissance, cette méthode n'a néanmoins pas encore été appliquée à des documents audios transcrits automatiquement et fortement bruités. Les méthodes issues des SVM ont également été utilisées couplées avec une représentation du document utilisant la fréquence des termes comme (Lan et al., 2005; Georgescul et al., 2006).

Une autre méthode, permettant la catégorisation des documents, est bâtie sur les hypothèses décrites dans la section 2.2.1 utilisant une règle de décision bayésienne ou une distance de Mahalanobis (Xing et al., 2002). Cette méthode est principalement utilisée dans le domaine de l'identification du locuteur. Ainsi (Dehak et al., 2011) proposent de mesurer la pertinence d'une nouvelle représentation d'un vecteur composé de mélanges de gaussiennes (Modèle de mélanges de gaussiennes ou *Gaussian Mixture Model* (GMM)), appelée *i*-vecteur, dans une tâche de reconnaissance du locuteur. De la même manière, (Bousquet et al., 2011) proposent d'évaluer la robustesse d'une représentation de la variabilité inter-session transformée à l'aide d'une matrice de covariance normalisée. La distance de Mahalanobis est souvent utilisée pour évaluer les performances des systèmes de vérification du locuteur.

La combinaison LDA-Mahalanobis n'a pas encore fait l'objet d'une étude approfondie durant une tâche de catégorisation de documents audios bruités. La section suivante décrit la manière dont une transcription d'un document audio est représentée

dans l'espace des mots (approche classique à base de TF-IDF) et dans un espace de thèmes (LDA).

2.2.3 Représentation d'un document audio transcrit automatiquement

Cette partie présente le système de catégorisation utilisant les mots discriminants extraits à partir de transcriptions très imparfaites. Le système est composé de deux parties principales. La première crée une représentation vectorielle des mots au moyen de deux approches non-supervisées : un vecteur de fréquences de mots Okapi/BM25 (Robertson, 2004) avec la méthode TF-IDF-Gini (Demiriz et al., 1999), et une représentation par espace de thèmes avec l'approche LDA (Blei et al., 2003). La seconde partie utilise les vecteurs extraits afin d'apprendre des classifieurs SVM ou un modèle bayésien. La figure 2.1 présente l'architecture globale du système de catégorisation proposé utilisant des transcriptions manuelles (Transcription manuelle d'un document (MAN)) et automatiques (Transcription provenant d'un Système de Reconnaissance Automatique de la Parole (RAP)).

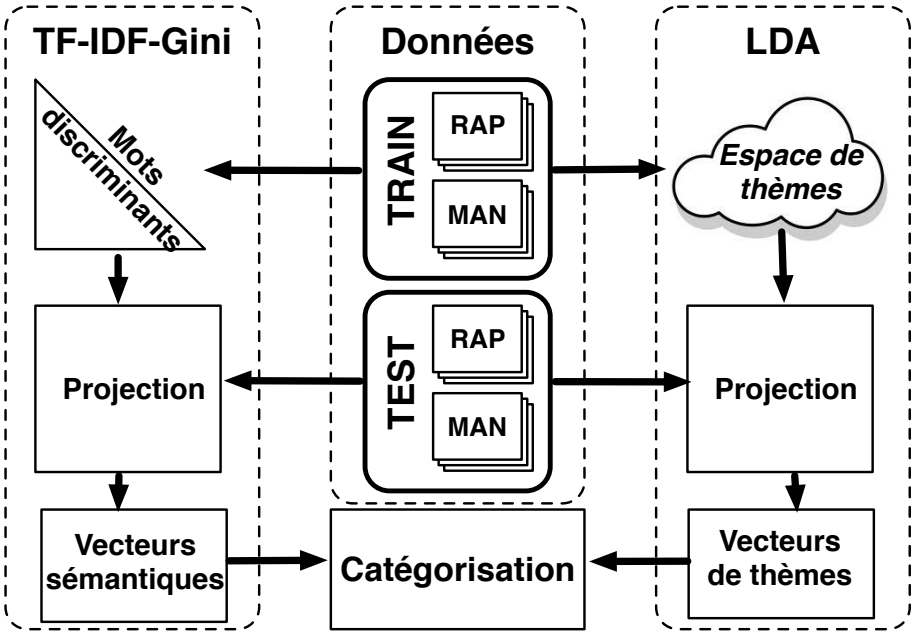


FIGURE 2.1 – *Architecture globale du système de catégorisation.*

2.2.3.1 Représentation par fréquence des mots

Considérons un corpus D de transcriptions d ayant un vocabulaire $\mathbf{V} = \{w_1, \ldots, w_N\}$ de taille N où d est vu comme un *sac-de-mots* (Salton, 1989b). Un mot w de $\mathbf{V}$ est choisi en fonction de son importance δ_t dans le thème t en calculant sa fréquence

(TF), sa fréquence inverse (IDF) (Robertson, 2004), et le critère de pureté de Gini (Demiriz et al., 1999) commun à tous les thèmes. Cet ensemble de scores δ compose le modèle de fréquence f :

$$\delta_t^w = tf_t(w) \times idf_t(w) \times gini_t(w)$$

Chaque catégorie $t \in \mathbf{T}$ possède son propre score $\delta_t^{|V|}$ et sa propre fréquence γ dans le modèle f :

$$\gamma_f^t = \frac{\#d \in t}{\#d \in D}$$

Ensuite, les mots ayant les scores les plus élevés Δ pour toutes les catégories $\mathbf{T}$ sont extraits, et constituent le sous-ensemble de mots discriminants $\mathbf{V}_\Delta$. Notons qu'un même mot w peut être présent dans différentes catégories, mais avec des scores différents (TF-IDF-Gini) normalisés (δ') en fonction de sa pertinence dans la catégorie :

$$\begin{aligned} \Delta(w) = P(w|f) &= \sum_{t \in \mathbf{T}} P(w|t)P(t|f) \\ &= \sum_{t \in \mathbf{T}} {\delta'}_t^w \times \gamma_f^t \\ &= \left\langle \overrightarrow{\delta'^w}, \overrightarrow{\gamma^f} \right\rangle \end{aligned} \tag{2.1}$$

Pour chaque transcription $d \in D$, un vecteur de caractéristiques sémantiques V_d^s est déterminé. La n^{eme} $(1 \leq n \leq |\mathbf{V}_\Delta|)$ caractéristique $V_d^s[n]$ est composée du nombre d'occurrences du mot w_n ($|w_n|$) dans d, et le score Δ de w_n (voir équation 2.1) dans l'ensemble des mots discriminants $\mathbf{V}_\Delta$:

$$V_d^s[n] = tf_t(w_n) \times \Delta(w_n) \tag{2.2}$$

2.2.3.2 Représentation du document dans un espace de thèmes

La représentation dans un espace de thèmes est réalisée au moyen de l'approche LDA. Un espace thématique m de K thèmes est obtenu avec, pour chaque thème z, la probabilité de chaque mot w de $\mathbf{V}$ sachant z $P(w|z)$, et la probabilité de chaque thème z $P(z)$.

Pour chaque document d du corpus D, un premier paramètre θ est défini en fonction d'une loi de Dirichlet de paramètre α. Un second paramètre ϕ est défini en fonction de la même loi de Dirichlet de paramètre β. Ensuite, pour générer chaque mot w du document d, un thème latent z est défini à partir d'une distribution multinomiale sur θ. Sachant ce thème z, la distribution des mots est une multinomiale de paramètre ϕ_z.

Le paramètre θ est défini pour tous les documents à partir du même paramètre initial α. Cela permet d'obtenir un paramètre reliant tous les documents ensemble (Blei et al., 2003). Plus de détails sur le modèle LDA sont donnés dans l'annexe D.

Projection des conversations/espace de thèmes : L'algorithme de *Gibbs Sampling* (Griffiths et Steyvers, 2002) est utilisé pour inférer un document d avec les K thèmes de l'espace thématique m. Cet algorithme s'appuie sur la méthode *Markov Chain Monte Carlo* (MCMC). Ainsi, le *Gibbs sampling* permet d'obtenir des échantillons de paramètres de distribution sachant un mot w d'un document de test et un thème donné z. Un vecteur de caractéristiques thématiques V_d^z de d est alors obtenu. La k^{me} caractéristique ($1 \leq k \leq n$) est la probabilité du thème z_k sachant le transcription d :

$$V_d^z[k] = P(z_k|d) \tag{2.3}$$

2.2.4 Méthodes de catégorisation

Cette section présente les deux méthodes de catégorisation utilisant la représentation vectorielle pour l'apprentissage d'un système de catégorisation SVM ou une approche gaussienne.

2.2.4.1 Classification à base de SVM

Durant cette étape, les systèmes de catégorisation sont entrainés à partir de la représentation vectorielle afin d'attribuer automatiquement le thème le plus pertinent à chaque conversation. Ce processus de catégorisation nécessite un classifieur multi-classes. La méthode *un-contre-un* est choisie avec un noyau linéaire. Cette méthode donne, en général, de meilleurs résultats que la méthode *un-contre-tous* (Yuan et al., 2012). Pour ce problème multi-catégories, T représente le nombre de catégories ou classes et $t_i, i = 1, \ldots, T$ représente les classes. Un système de catégorisation binaire (*un-contre-un*) est entraîné pour chaque paire de classes distinctes : tous les systèmes binaires $T(T-1)/2$ sont ensuite construits. Le système de catégorisation binaire $C_{i,j}$ est entrainé, t_i étant une classe positive et t_j une classe négative ($i \neq j$). Pour une nouvelle représentation vectorielle (vecteur de fréquence de mots, équation 2.2, ou vecteur d'espace de thèmes, équation 2.3) d'une transcription d du corpus de test, si $C_{i,j}$ signifie que d est dans la classe t_i, alors le vote pour la classe t_i est incrémenté de un. Sinon, le vote pour la classe t_j est augmenté de un. Une fois le vote de tous les systèmes de catégorisation achevé, la classe ayant le plus grand nombre de votes est attribuée à la transcription d.

2.2.4.2 Distance de Mahalanobis

Cette approche modélise le processus d'extraction des vecteurs de représentation, les considérant issus d'un modèle génératif. Les deux hypothèses de base sont relatives

à l'homoscédasticité (égalité entre les covariances des classes) du système de catégorisation (Petridis et Perantonis, 2004) :

- les distributions des classes au sein du corpus suivent une loi gaussienne,
- les covariances de chacune des classes sont égales.

Le système de catégorisation gaussien, s'appuyant sur un règle bayésienne, est combiné à une métrique d'évaluation permettant d'assigner une transcription automatique d à une catégorie ou classe la plus probable t.

Sachant un ensemble de transcriptions d'apprentissage, $\mathbf{W}$ représente la matrice de covariance intra-classes définie ainsi :

$$\mathbf{W} = \sum_{t=1}^{T} \frac{n_t}{n} \mathbf{W_t} = \frac{1}{n} \sum_{t=1}^{T} \sum_{i=0}^{n_t} \left(x_t^i - \overline{x_t} \right) \left(x_t^i - \overline{x_t} \right)^T \tag{2.4}$$

où $\mathbf{W_t}$ est la matrice de covariance de la t^{eme} catégorie C_t, n_t est le nombre d'exemples de la catégorie C_t, n est le nombre total de documents contenus dans le corpus d'apprentissage, $\overline{x_t}$ est la moyenne sur toutes les transcriptions x_t^i pour la t^{eme} catégorie, et T est le nombre total de catégories.

Toutes les transcriptions ne contribuent pas à la covariance totale de manière identique. Pour cette raison, le terme $\frac{n_t}{n}$ est introduit dans l'équation 2.4. L'algorithme 2 décrit les étapes nécessaires à l'estimation de la matrice de covariance.

Algorithm 2: Estimation de la matrice de covariance intra-classes $\mathbf{W}$.

Data: corpus D de documents contenus dans la matrice X
Result: matrice de covariance $\mathbf{W}$
for $t \in T$ **do**
  $C_t \leftarrow t^{\text{eme}}$catégorie
  $\mathbf{x}_t \rightarrow \overline{X[t]}$ ▷ moyenne sur toutes les transcriptions $\in C_t$
  $\mathbf{M} \rightarrow zero(K, K)$
  for $x \in C_t$ **do**
    $\mathbf{M} \leftarrow \mathbf{M} + (x - \mathbf{x}_t) \times (x - \mathbf{x}_t)^T$
  end
  $\mathbf{W} \leftarrow \mathbf{W} + \mathbf{M}$
end
$\mathbf{W} \leftarrow \frac{1}{n} \times \mathbf{W}$
return $\mathbf{W}$

Si les hypothèses d'homoscédasticité et de densité gaussienne du modèle conditionnel sont admises, une nouvelle observation (transcription contenue dans l'ensemble de test) peut alors être étiquetée comme appartenant à la catégorie la plus probable t_{Bayes} en utilisant le système de catégorisation gaussien s'appuyant sur la règle bayésienne

suivante :

$$\begin{aligned} t_{\text{Bayes}} &= \arg\max_t \ \mathcal{N}\left(x \mid \overline{x_t}, \mathbf{W}\right) \\ &= \arg\max_t \left\{ -\frac{1}{2}\left(x-\overline{x_t}\right)^T \mathbf{W}^{-1}\left(x-\overline{x_t}\right) + a_t \ , \right\} \end{aligned} \tag{2.5}$$

où $\overline{x_t}$ est le centroïde (moyenne) de la catégorie C_t, $\mathbf{W}$ représente la matrice de covariance intra-catégorie définie dans l'équation 2.4, $\mathcal{N}$ dénote la distribution normale, et a_t est la probabilité (logarithmique) de la catégorie C_t :

$$a_t = \log\left(P(C_t)\right) \tag{2.6}$$

Il est à noter que, avec ces hypothèses, cette approche bayésienne est similaire à l'approche géométrique de Fisher : x est assigné à la catégorie du centroïde le plus proche, selon la métrique de Mahalanobis (Xing et al., 2002) :

$$t_{\text{Bayes}} = \arg\max_t \left\{ -\frac{1}{2}||x-\overline{x_t}||^2_{\mathbf{W}^{-1}} + a_t \right\} \tag{2.7}$$

2.2.5 Protocole expérimental

Les expériences permettant d'évaluer l'impact de la représentation d'une transcriptions fortement bruitée sont réalisées en utilisant le corpus de conversations téléphoniques du projet DECODA (Bechet et al., 2012). Ce corpus est découpé en un ensemble de conversations pour la phase d'apprentissage des systèmes de catégorisation (*train*), et en un ensemble de validation (*test*) (voir tableau 2.1). Il est annoté manuellement et est composé de 8 catégories, comme indiqué dans le tableau 2.2.

Train	Test	**Total**
740	327	1067

TABLE 2.1 – *Corpus de conversations téléphoniques DECODA.*

L'ensemble d'apprentissage est utilisé pour définir un sous-ensemble de mots discriminants (voir partie 2.2.3.1). Ce sous-ensemble permet d'élaborer un ensemble de représentations fondées sur les caractéristiques basique TF-IDF-Gini. Dans ces expériences, le nombre de mots discriminants varie de 800 à 7 920 mots (nombre total de mots uniques contenus dans le corpus d'apprentissage). Le corpus de test contient 3 806 mots, 70,8 % d'entre-eux étant contenus dans le corpus d'apprentissage.

Un ensemble de 19 espaces de thèmes contenant un nombre de classes différent ($\{5, \ldots, 300\}$) est estimé au moyen de l'algorithme LDA à l'aide de l'outil Mallet (McCallum, 2002). De la même manière, un vecteur de probabilités sur les thèmes est calculé en projetant chaque dialogue du corpus dans chacun des 19 espaces de thèmes.

Catégories
Problèmes d'itinéraire
Objets trouvés
Horaires
Carte de transport
État du traffic
Tarifs
Procès verbaux
Offre spécial

TABLE 2.2 – *Ensemble de catégories composant le corpus de conversations téléphoniques DECODA.*

Ensuite, pour ces deux configurations, un classifieur SVM est entrainé au moyen de la librairie LIBSVM (Chang et Lin, 2011). Les paramètres sont optimisés par validation croisée sur le corpus de developpement.

Le système de RAP *Speeral* (Linarès et al., 2007) a été utilisé (voir annexe E pour de plus amples détails). Les paramètres des modèles acoustiques (230 000 gaussiennes / modélisation triphone) sont estimés au moyen d'une adaptation par maximum *a-posteriori* (MAP) à partir de 150 heures de parole (conditions téléphoniques). Un modèle de langage tri-gramme a été obtenu en adaptant un modèle de langage basique avec les transcriptions du corpus d'apprentissage de DECODA. Le vocabulaire contient 5 782 mots. Le taux d'erreur-mot (TEM) initial atteint 45,8 % sur le corpus d'apprentissage et 58,0 % sur le corpus de test. Ces TEM élevés sont principalement dus à la présence de nombreuses disfluences et à des conditions acoustiques bruitées, quand, par exemple, les utilisateurs appellent à partir de gares avec un téléphone portable. En filtrant références et transcriptions avec une liste de rejet (*stop-list*) de 126 mots-outils [2] on obtient au final un TEM de 33,8 % (apprentissage), et 49,5 % sur l'ensemble (test).

Les expériences sont menées au moyen des deux méthodes non-supervisées proposées (TF-IDF-Gini / LDA) sur les transcriptions manuelles (TMAN) et les transcriptions automatiques (TRAP) seules. Nous proposons également d'étudier la combinaison des transcriptions manuelles et automatiques (TMAN+TRAP) afin de voir si les erreurs de TRAP peuvent être compensées par les mots corrects (*i.e.* issus de la référence).

2.2.6 Résultats

Dans cette section, nous présentons les résultats permettant de répondre aux questions suivantes, relatives à la robustesse des représentations évaluées sur une tâche de catégorisation :

- Qu'apporte la représentation de haut niveau, produite par LDA, pour :
  — la catégorisation de dialogues en parole spontanée (robustesse au style de parole) ?

2. **http://code.google.com/p/stop-words/**

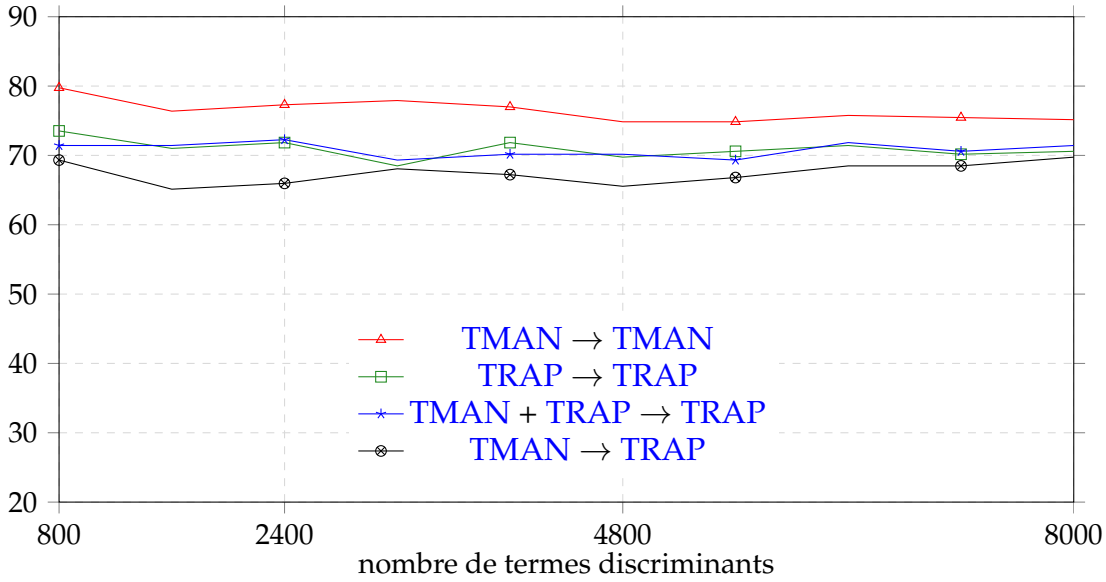


FIGURE 2.2 – *Performance en termes de précision (%) de la classification de thèmes en faisant varier le nombre de mots discriminants (TF-IDF-gini) en utilisant des SVMs.*

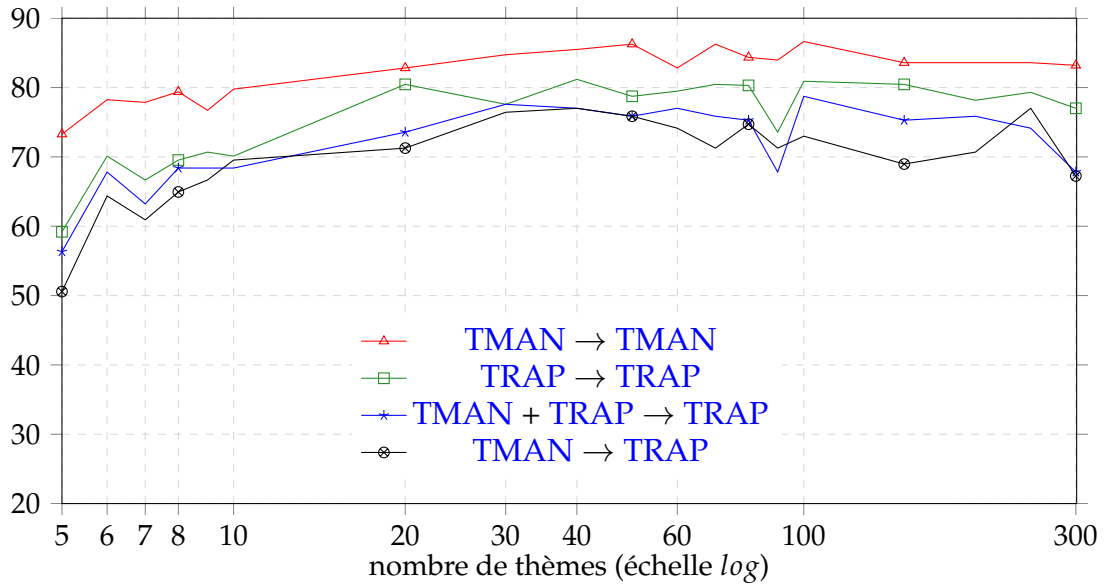


FIGURE 2.3 – *Performance en termes de précision (%) de la catégorisation de thèmes en faisant varier le nombre de dimensions de l'espace de thèmes (LDA) en utilisant des SVMs.*

— la catégorisation de dialogues mal transcrits (robustesse aux erreurs de transcription) ?

- Est-ce qu'une méthode de caractérisation tenant compte de la structure et des relations entre les composantes d'un vecteur de représentation obtient de meilleurs résultats qu'une méthode performante mais rompant cette relation

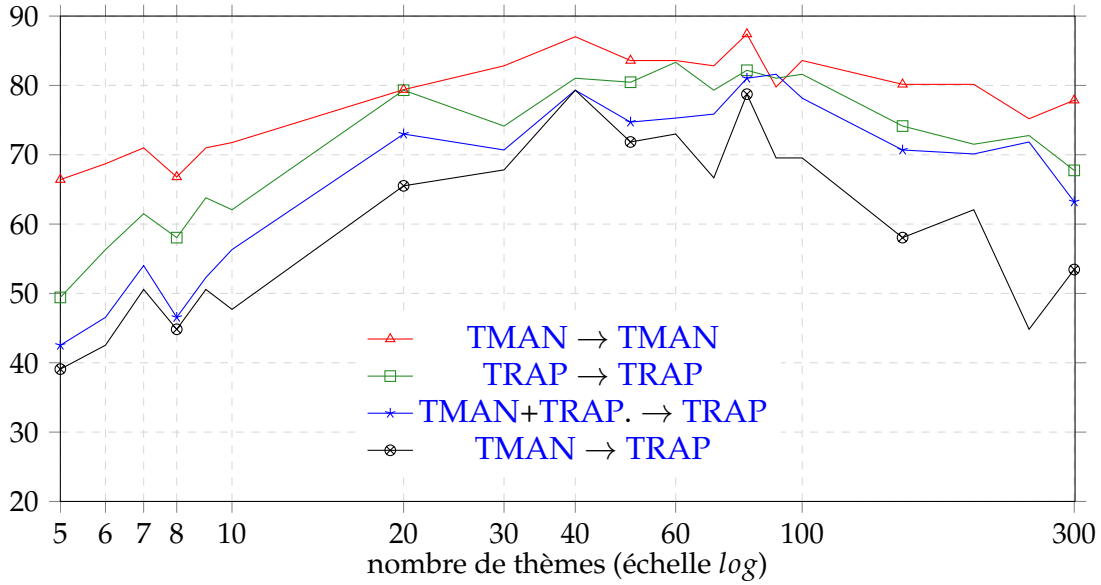


FIGURE 2.4 – *Performances observées en termes de précision (%) de la catégorisation de thèmes en faisant varier le nombre de dimensions de l'espace de thèmes (LDA) en utilisant la métrique de Mahalanobis.*

(section 2.2.6.2) ?

- Est-ce qu'une diminution de la dimension de l'espace de représentation permet une amélioration du nombre de transcriptions bien catégorisées (section 2.2.6.3) ?

- Enfin, quel est l'impact, sur la tâche de catégorisation, du choix des mots à prendre en considération lors du calcul du taux d'erreur-mot (section 2.2.6.4) ?

Les sections suivantes répondront successivement à ces quatre questions.

2.2.6.1 Performance de l'identification de catégories

Les figures 2.2 et 2.3 présentent les précisions de la classification de catégories obtenues avec des SVM par les approches TF-IDF-Gini et LDA sur le corpus de test pour les différentes configurations étudiées (sur des transcriptions manuelles ou automatiques, respectivement notées TMAN / TRAP) en faisant varier les conditions d'extraction des mots (nombre de mots discriminants et nombre de thèmes). Nous pouvons noter que la méthode LDA surpasse tous les résultats obtenus par l'approche TF-IDF-Gini (voir tableau 2.3).

Comme attendu, la configuration *TMAN* → *TMAN* donne les meilleurs résultats de classification avec un gain de 6,9 points avec la méthode LDA par rapport à une représentation basée sur le TF-IDF-Gini. Si nous comparons les configurations du corpus

Données		*Meilleure Précision (%)*			
Apprentissage	**Test**	**#mots**	**TF-IDF-Gini**	**#thèmes**	**LDA**
TMAN	TMAN	*800*	*79,7*	*100*	**86,6**
TMAN	TRAP	*8 000*	*69,7*	*40*	**77,0**
TRAP	TRAP	*800*	*73,5*	*60*	**81,4**
TMAN+TRAP	TRAP	*2 400*	*72,2*	*100*	**78,7**

TABLE 2.3 – *Précision de la classification de catégories en utilisant des SVM.*

d'apprentissage, nous notons également que les meilleurs performances sur le corpus de test TRAP sont obtenues avec le corpus d'apprentissage TRAP. Un gain de 10,9 points est constaté avec la méthode LDA en comparaison de l'approche TF-IDF-Gini sur les transcriptions automatiques. Il semble évident qu'utiliser des configurations d'apprentissage et de test comparables permet d'atteindre les meilleurs résultats de classification, et ce, peu importe que l'on traite des transcriptions manuelles ou automatiques.

Nous pouvons enfin noter que les performances obtenues avec l'approche LDA ont tendance à fluctuer lorsque le nombre de thèmes varie. Ceci peut s'expliquer par les taux d'erreur-mot (TEM) du corpus traité : en effet, les mots choisis comme *discriminants* dans des conditions particulières de l'espace de thèmes peuvent être mal transcrits dans des proportions élevées. Nous pouvons étayer cette remarque en analysant les résultats obtenus avec 90 thèmes (voir figure 2.3). Une baisse importante des performances est observée pour la condition d'apprentissage TRAP (TRAP $\rightarrow$ TRAP et TMAN $\rightarrow$ TRAP) alors qu'une faible baisse est constatée lorsque les transcriptions de références sont ajoutées dans le processus d'apprentissage (TMAN+TRAP $\rightarrow$ TRAP et TMAN $\rightarrow$ TMAN).

2.2.6.2 Impact des méthodes de catégorisation

Les figures 2.3 et 2.4 montrent les précisions de la catégorisation de transcriptions automatiques issues du corpus de test, représentées par des vecteurs issus de modèles LDA de taille (nombre de thèmes) différents, obtenue avec des SVM et une approche gaussienne pour toutes les configurations (TMAN/TRAP).

Nous pouvons constater que la méthode gaussienne obtient de meilleurs résultats que la méthode SVM, quel que soit les conditions de test (voir tableau 2.4). Comme il a été vu dans l'évaluation de la meilleure représentation du dialogue (section 2.2.6.1), la configuration TMAN $\rightarrow$ TMAN obtient le meilleur résultat en catégorisation, avec un gain de 0,8 point pour la distance de Mahalanobis.

Si l'on se concentre sur la configuration TRAP $\rightarrow$ TRAP, le système de catégorisation s'appuyant sur la règle bayésienne permet d'obtenir un gain de 1,9 points en comparaison de la méthode SVM.

Nous pouvons finalement remarquer que, si l'on considère un nombre de thèmes

inférieur à 100, la précision du système gaussien décroît. Cela peut s'expliquer par un nombre de données d'apprentissage trop faible pour permettre une estimation suffisamment précise de la matrice de covariance **W** ainsi que des paramètres du modèle LDA comportant un aussi grand nombre de thèmes.

Données		*Meilleure Précision (%) SVM*		*Meilleure Précision (%) gaussienne*	
Apprentissage	**Test**	*#thèmes*	**Test**	*#thèmes*	**Test**
TMAN	TMAN	*100*	86.6	*80/80*	**87.4**
TMAN	TRAP	*40*	77.0	*40/80*	**79.3**
TRAP	TRAP	*60*	81.4	*60/80*	**83.3**
TMAN+TRAP	TRAP	*100*	78.7	*90/80*	**81.6**

TABLE 2.4 – *Précision avec une méthode à base de SVM et une approche gaussienne.*

2.2.6.3 Impact de la réduction de l'espace de représentation par analyse en composantes principales

Les vecteurs de représentation des dialogues dans l'espace de thèmes LDA sont ici compactés pour mesurer l'impact d'une telle modification sur la structure de la distribution des thèmes au sein du document. Cette nouvelle représentation est obtenue par une analyse en composantes principales (ACP). La décomposition en composantes principales (ACP) ou *Principal Component Analysis* (PCA) est utilisée dans presque toutes les disciplines scientifiques. Introduite pour la première fois par (Cauchy, 1829; Pearson, 1901), sa formulation la plus récente est réalisée par (Hotelling, 1933) qui introduit le terme de composante principale. Les objectifs de l'ACP sont :

- extraire de l'information la plus importante depuis la table de données,
- compresser les données, en ne conservant que cette information jugée importante,
- expliquer et simplifier la description des données,
- analyser la structure des données observées et des variables.

L'ACP calcule de nouvelles variables appelées "composantes principales", obtenues comme une combinaison linéaire des variables originales. La première composante principale nécessite de représenter la variance la plus grande pour "expliquer" la plus grande partie de l'inertie totale de l'ensemble de données. Ensuite, la seconde composante est calculée sous la contrainte d'être orthogonale à la première et de composer la partie la plus grande de l'inertie restante. Les autres composantes sont calculées de la même manière. Les valeurs de ces nouvelles variables sont appelées "scores de facteurs" et sont interprétées géométriquement comme la "projection" des observations sur les composantes principales.

Celles-ci sont obtenues depuis une Décomposition en valeurs singulières ou *Singular Value Decomposition* (SVD) (voir annexe B) de l'ensemble de données $\mathbf{X}$, avec :

$$\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}^T \tag{2.8}$$

où $\mathbf{P}$ est une matrice de dimension $m \times l$ composée des vecteurs propres à gauche, $\mathbf{Q}$ est une matrice de dimension $n \times l$ composée des vecteurs propres à droite, et Δ est la matrice diagonale composée des valeurs singulières. l est le rang de la matrice $\mathbf{X}$ ($l \leq \min\{m, n\}$). La $m \times n$ matrice de scores de facteurs notée $\mathbf{F}$ est obtenue ainsi :

$$\mathbf{F} = \mathbf{P}\Delta \tag{2.9}$$

et peut être interprétée comme la matrice de projection puisque la multiplication de la matrice $\mathbf{X}$ et $\mathbf{Q}$ donne les valeurs de la projection des observations ($\mathbf{X}$) sur les composantes principales en utilisant l'équation 2.8 :

$$\begin{aligned} \mathbf{XQ} &= \left(\mathbf{P}\Delta\mathbf{Q}^T\right)\mathbf{Q} \\ &= \mathbf{P}\Delta \\ &= \mathbf{F} \end{aligned} \tag{2.10}$$

Plus d'informations sur l'ACP sont disponibles dans (Abdi et Williams, 2010).

Les figures 2.5 et 2.6 présentent les performances de catégorisation de conversations obtenues dans le contexte de transcription manuelle (TMAN → TMAN) en utilisant la distance de Mahalanobis.

La courbe de précision originale (en pointillée) présente les résultats obtenus avec la méthode fondée sur LDA utilisant l'espace de représentation originel (non-réduit), déjà présentée dans la figure 2.4 (TMAN → TMAN).

Ces résultats sont comparés avec ceux obtenus avec un espace de thèmes de taille réduite (courbe rouge avec triangles). Ces réductions sont réalisées depuis une ACP sur les représentations vectorielles des dialogues dans les espaces de thèmes LDA de taille plus grande que n ($n = 40$ dans la figure 2.5 ou $n = 80$ dans la figure 2.6). La dernière courbe (courbe bleue avec carrés) représente les résultats obtenus avec l'espace de thèmes de taille initiale n.

Considérons l'espace de thèmes de 80 dimensions présenté dans la figure 2.5. Nous pouvons voir que la précision obtenue avec la représentation originelle (courbe bleue) est d'environ 87,5 %. La précision obtenue après réduction de l'espace de représentation grâce à une ACP (LDA+ACP → taille de l'espace=40), est d'environ 85 %. Cette précision a été obtenue en réduisant le nombre de dimensions de 80 thèmes dans l'espace LDA à 40 valeurs dans l'espace ACP. Ainsi, nous pouvons noter que la précision décroît après réduction de l'espace de représentation par une ACP.

Pour toutes les autres dimensions ($n \neq 80$), la réduction de l'espace de thèmes de taille n améliore les résultats pour les cas ($n = 40$ ou $n = 80$). Cependant, nous pouvons

noter que la précision lors de la tâche de catégorisation ne dépasse jamais les résultats obtenus avec la représentation en espaces de thèmes issus de LDA (carré bleu).

Nous pouvons conclure que la réduction permet d'améliorer les résultats. Le fait d'avoir artificiellement augmenté le nombre de thèmes (granularité) sans augmenter le nombre de conversations dans le corpus d'apprentissage pour chacune des catégories, entraîne une baisse de la variabilité intra-catégories mais ne permet d'atteindre le score de précision optimale obtenu par l'espace de thèmes de taille initiale (n).

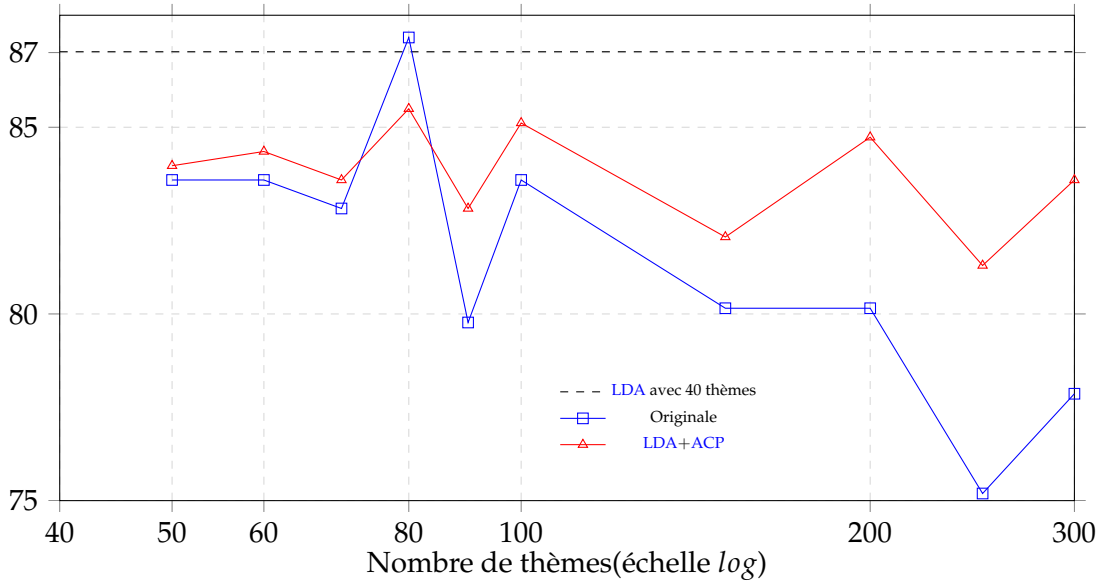


FIGURE 2.5 – *Précision en % avec les modèles d'espace de thèmes LDA m ($|m| > n$) pour $n = 40$.*

2.2.6.4 Précision de la transcription des mots discriminants

Alors que l'approche TF-IDF-Gini est clairement meilleure sur les transcriptions manuelles (voir tableau 5.2), les performances sont quasiment identiques sur les transcriptions manuelles et automatiques avec la méthode LDA (avec une précision respective de 86,6 % et 81,4 %). Nous pensons que l'approche LDA doit mieux gérer les erreurs contenues dans les transcriptions automatiques en choisissant les mots discriminants les mieux transcrits (*i.e.* ayant le plus faible TEM). Les figures 2.7-(a) et (b) compare les taux d'erreur-mot (TEM) des n mots discriminants extraits au moyen des méthodes TF-IDF-Gini et LDA sur toutes les configurations (TMAN / TRAP). Le score $s(w)$ est utilisé pour trouver les mots les plus pertinents de l'approche LDA selon la formule :

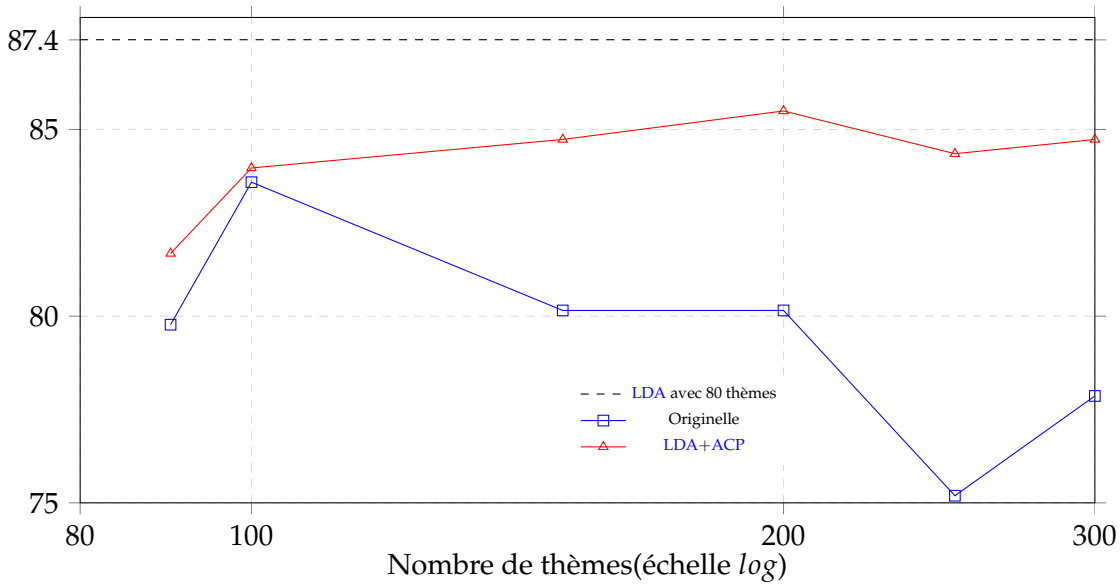


FIGURE 2.6 – *Précision en % avec les modèles d'espace de thèmes LDA m (|m| > n) pour n = 80.*

$$\begin{aligned} s(w) = P() &= \sum_z P(w|z)P(z) \\ &= \sum_z \theta_{w,z} \times \phi_z \\ &= \left\langle \overrightarrow{\theta_w}, \overrightarrow{\phi} \right\rangle \end{aligned}$$

où $\overrightarrow{\theta_w}$ est la représentation vectorielle d'un mot w dans tous les thèmes z de l'espace de thèmes, $\overrightarrow{\phi}$ est la représentation vectorielle de tous les thèmes z et $\langle \cdot, \cdot \rangle$ est le produit scalaire. Le TEM est ensuite calculé sur les mots les plus discriminants (poids de 1 pour chaque mot).

Si nous comparons tout d'abord les différentes configurations (TMAN / TRAP), nous pouvons noter que plus la précision de la classification est élevée (voir tableau 5.2), moins le TEM l'est. Ce constat est observé pour les deux méthodes. De plus, nous pouvons voir que le TEM obtenu avec l'approche LDA est légèrement plus bas que celui obtenu avec la méthode TF-IDF-Gini, peu importe la configuration considérée. Cela signifie que les mots discriminants extraits avec l'approche LDA sont mieux transcrits en comparaison de ceux obtenus avec la méthode TF-IDF-Gini, ce qui peut expliquer les meilleures performances de classification obtenues avec l'approche LDA.

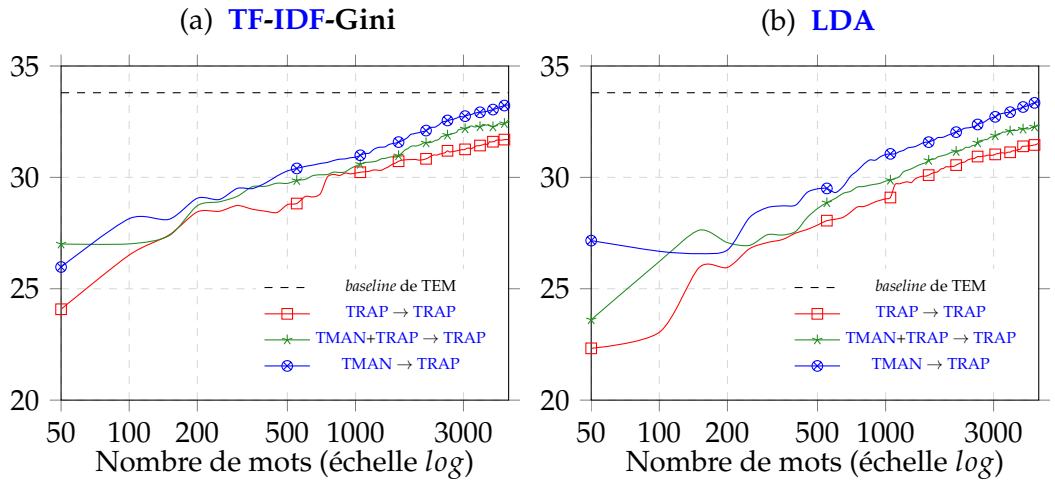


FIGURE 2.7 – *Taux d'erreur-mot en % des n mots discriminants extraits avec TF-IDF-Gini (a) et LDA (b).*

2.2.7 Conclusions pour la catégorisation de documents audios fortement bruités dans un espace de thèmes

Dans cette étude, nous avons présenté une architecture permettant d'identifier le thème d'une conversation en utilisant des transcriptions très imparfaites. Deux méthodes non-supervisées de représentation des conversations (TF-IDF-Gini et LDA) ont été comparées. Nous avons montré que la représentation par espace de thèmes obtenue avec la méthode LDA surpasse les résultats de classification obtenus avec la représentation classique TF-IDF-Gini. La précision de la classification atteint 86,6 % sur les transcriptions manuelles et 81,4 % sur les transcriptions automatiques, avec un gain respectif de 6,9 et 10,9 points.

La seconde partie de cette étude s'est concentrée sur le choix de la meilleure méthode de classification. Ainsi, nous avons démontré que les intuitions à propos de la gaussianité des distributions au sein des catégories et l'égalité des covariances des classes discutées dans cette étude sont pertinentes. La représentation en espace de thèmes combinée à l'approche gaussienne obtient de meilleurs résultats que ceux obtenus en utilisant l'approche à base de SVM. La précision, en termes de transcriptions bien catégorisées, atteint 87,4 % pour des transcriptions manuelles et 84,4 % avec des transcriptions automatiques (TRAP), avec des gains respectifs de $0,8$ et $1,9$ points.

Dans la troisième partie de cette étude, nous avons montré que la réduction de l'espace (LDA+ACP) améliore les résultats obtenus avec la représentation thématique originale de la transcription. Cependant, nous notons également que cette réduction ne permet pas d'atteindre les résultats obtenus par la représentation thématique sans réduction.

Nous avons également discuté du lien possible entre performance de classification et précision de la transcription. L'analyse proposée a montré que les meilleurs résultats

de classification sont obtenus avec des configurations extrayant les mots discriminants ayant les taux d'erreur-mot les plus faibles. Ces résultats prometteurs conduiront à une analyse qualitative plus détaillée dans des travaux futurs. En effet, cette étude préliminaire pourrait être fortement étendue avec de nouvelles analyses, en prenant par exemple en compte le poids des mots discriminants dans l'évaluation de la précision des transcriptions. Une perspective générale serait de proposer une solution pour estimer les performances de classification selon la qualité des transcriptions. Dans un contexte lié aux métriques d'évaluation, il serait intéressant de trouver une façon d'estimer la précision des transcriptions automatiques sur des tâches spécifiques, le taux d'erreur-mot n'étant pas le meilleur indicateur de la qualité d'une transcription dans un cadre applicatif.

2.3 Représentations fondées sur des espaces de thèmes LDA dans diverses tâches de RI

La section précédente a évalué l'apport d'une représentation s'appuyant sur un espace de thèmes issu de LDA pour une catégorisation de transcriptions automatiques fortement bruitées. Cette représentation du document a également été évaluée dans des contextes applicatifs de recherche d'information. La présente section décrit brièvement les méthodes proposées, les tâches sur lesquelles elles ont été évaluées ainsi que les résultats obtenus.

2.3.1 Contextualisation d'un message court dans un espace de thèmes

Les plateformes de microblogging sont des espaces d'échange destinés à la communication rapide entre internautes. Ces sites de partage se développent à la même vitesse que la masse de données disponible sur le Web et offrent aux utilisateurs une façon simple et ludique de disséminer des idées, des opinions ou des faits communs propres à la vie quotidienne. Cette communication utilise des formes contraintes, souvent des messages courts. Par exemple, le service *Twitter* ne permet pas l'envoi de messages dépassant la taille de 140 caractères. Cette contrainte conduit l'utilisateur à employer un vocabulaire souvent inhabituel, bruité, comportant un certain nombre de termes nouveaux, parfois mal orthographiés ou volontairement transformés (Choudhury et al., 2007).

L'objectif de ces messages est de partager le maximum d'information dans des structures grammaticales comportant le moins de caractères possible. Il peut être alors difficile de comprendre le message véhiculé par un Message court ou *Short Text Message* (STM) sans une connaissance du contexte général dans lequel il a été généré.

Pour contourner cette limitation en taille, nous proposons de représenter le STM dans un espace de plus haut-niveau que l'espace de représentation lexicale (Morchid et al., 2013b). Cette représentation thématique peut être vue comme une expansion du message court permettant d'améliorer la caractérisation de ce message. Cette approche

s'appuie sur une représentation thématique utilisant LDA : le STM est projeté dans un espace estimé depuis un grand corpus de documents, permettant d'identifier les thèmes latents associés au message. En résultat de cette projection, le message court est représenté par l'ensemble des termes qui le composent ainsi que des thèmes issus de LDA qui lui sont le plus proches, dans l'objectif d'améliorer la compréhension de celui-ci.

Afin de mettre en lumière l'intérêt d'une telle approche, nous évaluons cette méthode de représentation d'un message court, fondée sur des espace de thèmes LDA, dans le contexte de la campagne INEX 2012 (SanJuan et al., 2012).

L'objectif de cette campagne est de rechercher le contexte associé à un message court (ou *tweet*), issu de *Twitter*, dans l'idée d'aider l'utilisateur qui lit ce message court à comprendre plus aisément le message véhiculé. Nous proposons d'extraire un ensemble de mots-clés qui seront utilisés par des systèmes de recherche d'information (RI) et de résumés automatiques (RA) fournis par les organisateurs de la campagne INEX 2012 afin de déterminer le contexte le plus proche du message véhiculé par un *tweet*. La liste de mots-clés est composée par les termes contenus dans le *tweet* et par des termes extraits depuis l'ensemble des thèmes issus de LDA, proches du *tweet*.

Ce système s'est classé entre la 7^{eme} et la 12^{eme} place lors de la campagne INEX 2012 pour un total de 33 participants. La performance de notre système montre que cette approche permet une bonne représentation de haut-niveau d'un message court (STM). Cette tâche est d'autant plus complexe que les messages issus de *Twitter* sont composés d'un vocabulaire peu standard, qu'ils contiennent de nombreuses erreurs de grammaire et des termes nouveaux, parfois inventés pour contourner la limite de 140 caractères.

Au delà du cadre de la compétition, ce travail montre ce que les espaces thématiques peuvent apporter en termes de robustesse à des formes d'écritures très atypiques.

2.3.2 Extraction de mots-clés dans des transcriptions de vidéos communautaires

Les plates-formes de partage de vidéos sur Internet se sont fortement développées ces dernières années. En 2011, YouTube augmentait d'une heure d'enregistrement déposée toutes les secondes. Malheureusement, l'utilisation de ces collections de vidéos souffre de l'absence d'informations structurées et fiables. L'indexation réalisée par l'hébergeur repose essentiellement sur les mots-clés fournis par les utilisateurs, éventuellement sur les résumés ou le titre des documents. Malheureusement, ces méta-données sont souvent incomplètes ou erronées, parfois volontairement : certains utilisateurs choisissent des mots-clés qui favorisent le référencement jusqu'à s'éloigner significativement du contenu réel de la vidéo déposée. Ceci implique donc des *tags* non représentatifs du contenu même de la vidéo.

Un des problèmes majeurs de cet enchaînement extraction/analyse des contenus est lié au composant de reconnaissance de la parole, qui est souvent peu performant

sur des données Web, dont la diversité de forme et de fond est extrême et qui sont généralement éloignés des conditions d'entraînement des systèmes.

Une solution, permettant d'améliorer la tolérance du système d'analyse aux erreurs de reconnaissance, est d'utiliser les représentations thématiques décrites dans les sections précédentes. Elles reposent sur l'idée que le niveau lexical est particulièrement sensible aux erreurs de reconnaissance et qu'une représentation de plus haut niveau pourrait permettre de limiter l'impact négatif de ces erreurs sur les modules d'analyse. Le document source est projeté dans un espace thématique dans lequel le document peut être vu comme une association de thèmes. Cette représentation intermédiaire est obtenue par une analyse latente de Dirichlet (LDA) appliquée sur un grand corpus de textes.

La méthode d'étiquetage automatique de vidéos que nous proposons utilise cette décomposition pour déterminer les mots-clés caractéristiques du document source (Morchid et Linarès, 2013a).

Evalué sur un ensemble de réduit de 100 vidéos transcrites avec le système de reconnaissance de la parole du LIA, les premiers résultats que nous avons obtenus montrent à la fois un gain constaté avec la méthode utilisant les espaces de thèmes LDA, et des résultats particulièrement bas dans l'absolu. La meilleure configuration atteint 5,5 % en termes de précision, alors que la méthode classique atteint elle 2,6 %. Cependant les *tags* de l'utilisateur ne sont pas de "parfaites" références (autant que ces références puissent exister) et sont, parfois, probablement impossibles à prédire, car dépendantes du point de vue de l'utilisateur, de sa culture, de ses intentions etc. La méthode proposée semble néanmoins bien plus performante que des méthodes classiques d'extraction de mots-clés par estimation de fréquences relatives (Morchid et Linarès, 2013a).

2.3.3 Catégorisation de messages courts représentés dans un espace de thèmes pour la prédiction du *Buzz*

La tâche de catégorisation, vue dans la section précédente, est appliquée dans cette partie à des messages courts (STM) afin de déterminer si un STM sera fortement relayé ou non. Le pouvoir de dissémination d'une information dans une structure aussi grande et aussi interactive qu'est Internet, est susceptible d'intéresser le simple utilisateur, le journaliste, ou le politicien désireux d'évaluer la portée d'une information ou d'une opinion. De façon plus générale, le modèle économique d'Internet s'appuie sur la fréquentation, et donc de la capacité d'anticiper la popularité d'un intérêt majeur.

Les informations diffusées sur les plateformes de microblogging sont de portées très variables, mais l'audience potentielle et la rapidité du support favorisent la médiatisation "explosive" de certaines d'entre-elles. Cette étude traite de la prédiction *a priori* de ces "explosions d'activité médiatique" (Froissar, 2007) que l'on nommera *buzz*. La prédiction des *buzz* est une tâche difficile notamment parce que le phénomène est dépendant de paramètres très divers, liés à l'événement, à ses conséquences éventuelles, à la sensibilité du public, etc. mais aussi aux aspects dynamiques de la médiatisation :

les canaux par lesquels l'information circule, les relais, la tendance du bruit médiatique à s'auto-alimenter, ...Ces difficultés se trouvent augmentées par la dimension du Web, et la dispersion et la fragmentation des informations qui s'y trouvent. Plusieurs études, portées sur les modèles de diffusion de l'information (Bass, 2004; Goldenberg et al., 2001). (Bass, 2004; Goldenberg et al., 2001; Kempe et al., 2003), étudient l'impact du "bouche-à-oreille" et du processus de "publicité virale" pour la diffusion de l'information. D'autres études ont porté sur la dispersion de l'information en utilisant des modèles bâtis sur le seuillage (Kempe et al., 2003). De ce point de vue, Twitter est un espace d'expérimentation plus facile à traiter que le Web dans sa globalité.

Nous présentons dans (Morchid et al., 2014c) une méthode de prédiction des *buzz* appliquée à la prédiction des pics de ré-émissions (*retweets*) des messages postés sur Twitter. Certaines techniques de prédiction de messages, spécifiquement sur Twitter, ont fait l'objet d'études (Romero et al., 2011; Yang et Counts, 2010). Concrètement, nous considérerons qu'un *tweet* [3] aura fait du *buzz* si son nombre de *retweets* [4] dépasse un seuil fixé *a priori*.

La prédiction des *buzz* est réalisée par l'analyse du contenu textuel des STM. Nous proposons d'extraire de ces contenus trois types d'indicateurs dont nous pensons qu'ils participent à la probabilité de rediffusion d'un message : la popularité des termes du *tweet*, la saillance thématique, et l'expressivité. Pour chacun d'eux, nous proposons des caractéristiques qui sont utilisées comme variables d'entrée d'un prédicteur neuronal (réseau de neurones).

Les expérimentations montrent que les résultats obtenus en utilisant l'extraction de mots-clés par LDA, sont meilleurs que ceux issus d'un simple TF-IDF-Position Relative ou *Relative Position* (RP), ce qui confirme l'idée qui avait motivé cette approche : le passage par cette représentation intermédiaire améliore la robustesse du système à la langue "bruitée" de Twitter.

2.4 Conclusion générale du chapitre

L'apport d'une représentation dans un espace de thèmes issu d'une analyse latente de Dirichlet (LDA), a montré son efficacité dans différents problèmes auxquels le domaine de la recherche d'information est confronté. Le présent chapitre a présenté certaines applications tirant un profit certain d'une telle représentation de haut-niveau.

Les idées qui ont initialement motivé le travail reporté dans ce chapitre sont :

— La projection des documents bruités dans un espace thématique a un effet "filtrant" de la forme de surface. En effet, l'effet filtrant attendu repose sur l'idée que le bruit au niveau lexical n'est pas assez cohérent pour se projeter au niveau

3. Un *tweet* est un message diffusé à un ensemble de personnes, appelées abonnés ou *followers*, qui suivent l'activité d'un individu sur la plateforme.

4. Le *retweet* est le mécanisme de ré-émission permettant à un utilisateur de renvoyer un message vers ses propres abonnés.

sémantique. Pour cette raison, un espace de thèmes est moins enclin à contenir ce bruit au sein des thèmes qui le composent.

— L'espace thématique est plus proche des niveaux d'interprétation (thèmes) qui correspondent aux tâches finales d'analyse des contenus.

Nos expérimentations sur une grande diversité de tâches confirment cette idée. Elles ont aussi montré des gains en robustesse sur différents types de bruit (style de parole, transcriptions imparfaites, style d'écriture, ...). Un certain nombre de questions restent cependant ouvertes, en particulier sur la granularité du modèle et le fait qu'il repose sur des représentations en sac-de-mots dans lesquelles la structure temporelle des documents est perdue. Ces questions sont abordées dans les chapitres suivants.

Deuxième partie

Multiples représentations thématiques de documents bruités pour une catégorisation robuste

Chapitre 3

Projection d'un document bruité dans des espaces multiples

Sommaire

Résumé

Nous avons montré l'intérêt de la représentation thématique de documents bruités dans diverses tâches de catégorisation ou de recherche d'information mais cette méthode souffre néanmoins de plusieurs faiblesses, et notamment d'une granularité définie a priori. Ce choix fixé, la représentation d'un document dans l'espace de thèmes reflétera uniquement une vision partielle des contenus, liée à la granularité choisie. Ce chapitre présente une méthode de représentation multi-granulaires d'un document ; elle est évaluée d'abord sur une tâche de classification d'événements sociaux dans le cadre de la campagne d'évaluation MediaEval 2011, puis sur un problème de catégorisation de documents transcrits automatiquement.

3.1 Introduction

Le chapitre 2 montre l'apport d'une représentation thématique pour différentes tâches de recherche d'information et, plus particulièrement, pour la tâche de catégorisation de documents bruités.

Une difficulté majeure d'une telle représentation est la spécification *a priori* des caractéristiques de ces espaces abstraits. La qualité du modèle LDA, décrit dans l'annexe D, dépend fortement du choix des hyper-paramètres du modèle, ainsi que du nombre de thèmes le composant. Ainsi, les figures D.1, D.2 et D.3 montrent l'influence du paramètre α sur la distribution des thèmes au sein du corpus d'apprentissage du modèle LDA. De plus, le choix de la valeur de α impact la distribution des thèmes associée à un document n'appartenant pas à l'ensemble d'apprentissage du modèle. Ainsi, une valeur de α faible, permettra d'assigner un nombre réduit de classes au document et, inversement, une valeur élevée (proche de 1) de α impliquera une distribution plus uniforme des classes sachant le document d. Ceci est illustré dans la figure D.3.

Une caractéristique déterminante lors de l'élaboration d'un espace de thèmes s'appuyant sur LDA est le choix du nombre de thèmes. Le "bon" choix du nombre de classes (ou thèmes) est une étape cruciale, spécialement quand les documents composant le corpus traitent de plusieurs sujets. Plusieurs études essaient de trouver une méthode pertinente et efficace permettant de résoudre le problème de dimensionnement du modèle LDA. Dans (Arun et al., 2010), les auteurs proposent d'utiliser une décomposition en valeurs singulières (SVD, voir annexe B) pour représenter les frontières entre les termes contenus dans le vocabulaire. Ensuite, si les valeurs singulières de la matrice terme-thème $\mathbf{M}$ sont égales à la norme des lignes de $\mathbf{M}$, ceci indique alors que le vocabulaire est distribué de manière homogène dans les thèmes composant l'espace. Cependant, ce processus est très coûteux en termes de temps de traitement sachant que la masse de documents à traiter devient de plus en plus grande depuis l'avènement du Web.

(Teh et al., 2004) proposent une méthode s'appuyant sur le Processus Hiérarchique de Dirichlet ou *Hierarchical Dirichlet Process* (HDP) pour estimer le "bon" nombre de thèmes, en réalisant l'hypothèse que les thèmes composant l'espace thématiques sont structurés hiérarchiquement. Le modèle HDP est comparé à celui issu de LDA sur un ensemble de données identiques dans (Zavitsanos et al., 2008). Les auteurs y présentent une méthode d'apprentissage de la bonne profondeur de la structure hiérarchique d'une ontologie sachant le nombre de thèmes composant le modèle LDA.

L'étude présentée par (Cao et al., 2009) est très similaire. Les auteurs déterminent le "bon" nombre de thèmes composant le modèle, en utilisant la corrélation moyenne entre chacune des paires de classes à chaque niveau hiérarchique de la structure.

Toutes ces méthodes font l'hypothèse que le document ne peut avoir qu'une seule représentation et que le problème est de trouver le nombre optimal de classes.

Une alternative envisageable consiste à utiliser plusieurs représentations d'un document dans des espaces de thèmes de tailles diverses. Cette représentation multiple

pourrait permettre de représenter différents aspects du document selon la granularité du modèle : un espace contenant peu de thèmes aura tendance à contenir des classes très larges, portant sur des sujets très généraux, alors qu'un espace de grande dimension comportera des thèmes moins larges et plus précis. Nous pouvons voir que différentes granularités apportent différents points de vue sur le document et que ces vues peuvent être complémentaires. L'idée, ici, est de multiplier les vues de façon à ne pas effacer trop tôt une information qui pourrait se révéler utile.

Cette question de la granularité des modèles a été discutée par divers auteurs et a donné lieu à des propositions diverses. (Phan et al., 2008) étudient l'apport d'une représentation multiple de messages courts lors d'une tâche de classification. Ces messages courts prennent la forme de résumés d'articles ou de pages Web appelés *snippet*. Dans (Chen et al., 2011), les auteurs comparent plusieurs configurations en termes de taille de modèle (nombre de classes contenues dans l'espace de thèmes LDA) et de méthode de catégorisation (SVM et Maximum d'entropie). Cette représentation multiple d'un message court est enfin combinée pour fournir un nouvel ensemble de caractéristiques du message. La combinaison résulte simplement de la concaténation des caractéristiques individuelles de chacune des représentations dans chaque espace de thèmes pondéré par un poids μ. Cette nouvelle représentation multi-thèmes d'un message court obtient de meilleurs résultats (85,31%) que les représentations dans chacun des espaces de thèmes pris séparément (81,58%) pour les deux méthodes de catégorisation.

Le modèle proposé par (Chen et al., 2011) combine des espaces de thèmes pour constituer un ensemble de caractéristiques permettant de représenter efficacement un document de taille réduite. Dans (Titov et McDonald, 2008), les auteurs proposent de construire un espace de thèmes approprié à la tâche d'analyse d'opinions. Une représentation classique fondée sur un espace de thèmes issu de LDA ou PLSA ne peut alors pas convenir. En effet, ces modèles reposent sur le modèle en *sac-de-mots* (*bag-of-words*), permettant uniquement de modéliser les co-occurrences de termes au sein de chaque document composant le corpus d'apprentissage du modèle. Ceci est convenable tant que l'objectif d'utilisation d'un tel modèle est de trouver un ensemble de thèmes proches d'un document donné. Cette représentation fournit alors, dans le contexte d'analyse d'opinions, une vision globale du document en généralisant les thèmes que celui-ci aborde (par exemple *hôtel en France*, *auberge de jeunesse*, ...), mais ne permet pas de déterminer un ensemble de concepts liés à l'opinion de l'utilisateur (par exemple *propre*, *spacieux*, ...). Ainsi, (Titov et McDonald, 2008) introduisent un nouveau modèle génératif probabiliste nommé MG-LDA permettant de modéliser aussi bien des concepts *généraux* (par exemple *hôtel*, *musique*, *film*, ...) que des concepts *locaux* (par exemple *propre*, *nourriture*, *action*, ...). Les auteurs étendent le modèle LDA en introduisant une fenêtre glissante permettant de localiser, au niveau de la phrase, le thème associé à un terme du document. Ainsi, un terme permet de changer le thème global (via une allocation de Dirichlet classique au niveau du document) et au niveau local (via une estimation du thème local au niveau de la phrase contenue dans la fenêtre glissante). Ce modèle permet un gain substantiel de 2 % en termes de prédiction d'opinion (Snyder et Barzilay, 2007).

Cependant, une représentation fondée sur un ensemble d'espaces de thèmes et non un seul, présente une faiblesse liée à la relation entre les classes d'un même modèle. Dans (Blei et Lafferty, 2006a), les auteurs montrent que les thèmes contenus dans un espace de thèmes issu de LDA sont liés. De plus, les auteurs dans (Li et McCallum, 2006) considèrent une classe comme un noeud d'un graphe acyclique et comme une distribution sur les autres classes composant le même espace de thèmes.

Pour évaluer la pertinence d'une représentation d'un document bruité dans des espaces thématiques multiples, deux études sont réalisées sur des documents décrivant des images postées sur un célèbre site de partage, et des documents issus d'un système de reconnaissance automatique de la parole. Ces évaluations sont réalisées sur deux corpus distincts et font appel à des approches différentes.

Dans la section 3.2 sont présentées deux études portant à mesurer la pertinence d'une représentation multiple lors d'une tâche de catégorisation de documents bruités venant de sources identiques à celles présentées dans le chapitre 2. Nous nous intéresserons donc au Web, avec la catégorisation d'images issues de Flickr dans la section 3.2.1, ainsi qu'à des transcriptions imparfaites issues de conversations agent/client dans la section 3.2.2.

3.2 Contributions

Nous avons montré l'efficacité d'une représentation dans un espace de thèmes LDA pour le traitement des documents fortement bruités issus du Web ou de transcriptions automatiques de documents parlés (voir le chapitre 2). Nous utiliserons donc dans ce chapitre des corpus analogues au chapitre 2 pour évaluer la pertinence d'une approche multi-vue fondée sur la projection du document dans un ensemble d'espaces de thèmes de granularités différentes.

La première étude, décrite dans la section 3.2.1, porte sur la détection d'événements sociaux dans un ensemble d'images issues d'un site de partage. La section 3.2.2 présente une étude visant à trouver le thème principal d'une conversation entre un agent et un utilisateur, en utilisant une représentation thématique dépendante des locuteurs.

3.2.1 Détection d'événements sociaux dans des documents bruités issus du Web

Les sites de partage d'images, comme *Picasa*, *Flickr* ou *Drawin*, permettent à l'utilisateur de partager facilement et rapidement des photos ou de naviguer au sein de galleries d'images. Cependant, la recherche d'information dans des bases de données de si grandes tailles peut s'avérer très difficile et très coûteuse en termes de temps de traitement du serveur d'indexation. La manière classique de référencer un tel nombre de documents est d'utiliser les méta-données (ou *tags*) représentant l'image comme des caractéristiques (fréquence de mots, ...) de l'image. Dans un monde parfait, les *tags*

devraient être choisis par un expert, ce qui est très coûteux pour une plateforme de partage gratuite contenant des millions d'images.

Par conséquent, les plateformes de partage d'images abandonnent à l'utilisateur l'annotation de ses propres images. Cette vision participative de l'annotation entraîne inévitablement des erreurs (fautes d'orthographe, mots hors-vocabulaire, étiquettes manquantes, ...). Dans une telle situation, un système classique d'indexation d'images utilisant la fréquence des termes représentée par les *tags*, ne peut répondre efficacement à une requête donnée.

Nous présentons ici une représentation au moyen de modèles thématiques d'images étiquetées, appliquée à une tâche de détection d'événements sociaux. Cette tâche est une partie du projet *Topic Detection and Tracking* (TDT) (Allan et al., 1998b). L'objectif est de détecter un événement social se déroulant (ou allant se dérouler) dans un lieu précis ou à une date précise (Allan et al., 1998a). Un des premiers travaux proposés dans la recherche d'événements a été réalisé par (Yang et al., 1998). Les auteurs ont utilisé un algorithme de classification pour localiser un événement dans un grand corpus de données. Dans (Golder et Huberman, 2007), les dépendances entre *Flickr* et *Last.fm* sont étudiées en utilisant un ensemble d'étiquettes extraites depuis *Del.icio.us*. Les auteurs dans (Rattenbury et al., 2007) essaient d'extraire des contenus sémantiques depuis les méta-données associées à des images postées sur Flickr. Ces travaux se sont heurtés à deux verrous majeurs, liés au paradigme de modélisation d'un événement et à la quantité de données nécessaire à l'apprentissage d'un système robuste de détection s'appuyant sur une approche statistique.

Cette représentation doit permettre la détection automatique d'événements sociaux depuis un ensemble d'images issues de la plateforme de partage Flickr, en utilisant uniquement le contenu textuel composé des méta-données des images. Ce système repose sur une représentation multiple du contenu textuel d'une image, ainsi que sur une stratégie faiblement supervisée permettant d'estimer un ensemble de modèles depuis un ensemble de données partiellement annotées.

Comme indiqué dans l'annexe D, LDA est un modèle statistique considérant que le document est composé d'un mélange de thèmes cachés. Ces concepts, appartenant à un document, sont liés par une variable latente qui est la distribution des termes composant le document. Le choix du nombre de thèmes composant l'espace de représentation est donc une donnée cruciale pour permettre en même temps :

- une bonne distribution des termes au sein des classes, et des classes au sein des documents composant le corpus d'apprentissage,
- une bonne généralisation à des documents non rencontrés lors de cette phase d'apprentissage.

En effet, la granularité du modèle dépend directement de la taille du modèle : un nombre restreint de classes entraîne une représentation grossière des classes qui peuvent alors être vues comme des domaines ou bien des thèmes. D'un autre côté, un espace thématique contenant un nombre élevé de classes contiendra des entités pouvant être assimilées à des thèmes précis ou concepts. Le choix de cette granularité dépend alors des objectifs et de la tâche pour laquelle le modèle est réalisé.

Les travaux précédents sur la granularité des modèles LDA se sont généralement focalisés sur représentation dite "optimale" d'un document. Une représentation unique d'un document bruité, comme une page Web ou un *tag* relatif à une image, ne peut utiliser de telles méthodes s'appuyant sur le contenu sémantique du document. De plus, la granularité d'un modèle thématique étant dépendante du type de document traité, les documents bruités fortement hétérogènes rencontrés dans ce genre de tâche ne peuvent être représentés dans un espace unifié pour tous les documents. L'approche présentée dans cette étude, considère un ensemble d'espaces de thèmes de granularités différentes comme des vues complémentaires d'un même document. Cette combinaison doit permettre d'améliorer la détection d'événements sociaux en dépit de la variabilité liée à la représentation multiple.

Un autre point clé est lié au nombre de données nécessaire pour une bonne estimation des modèles statistiques pouvant détecter l'événement social. Nous proposons une approche faiblement supervisée pour estimer la signature d'un événement ; cette méthode implique conjointement une annotation humaine d'un nombre réduit de données et un grand nombre de documents collectés depuis le Web qui sont probablement (mais pas obligatoirement) issus des classes attendues.

La section suivante présente l'architecture du système proposé. La section 3.2.1.2 décrit le protocole expérimental de la campagne SED de MediaEval 2011. Les résultats sont rapportés dans la section 3.2.1.3 avant de conclure dans la section 3.2.1.4.

3.2.1.1 Système de détection d'événements fondé sur une représentation multi-granulaires

Cette section présente le système d'extraction d'un ensemble de photos pertinentes répondant au mieux à une requête donnée. Ce système s'appuie sur une représentation multiple (différentes granularités) issue de LDA.

Approche générale

Le système de détection d'événements sociaux est composé de deux modules appliqués successivement. Le premier module (*WEB*) consiste en l'extraction d'un ensemble de pages Web depuis une requête, dans le but d'estimer un modèle fondé sur la fréquence de mots. Ce modèle est ensuite utilisé pour classifier des images en deux catégories (*pertinent* et *non-pertinent*), en fonction d'un seuil.

Le premier module suit une stratégie classique de recherche d'image, en comparant la fréquence des termes (*tags* dans notre cas) composant celle-ci avec la fréquence des termes composant le modèle *WEB*.

Le second module (*SVM*) a pour objectif de trouver les images pertinentes sachant la requête, au milieu de celles considérées comme non-pertinentes par le premier module (*WEB*). Ce processus prend en entrée un ensemble d'images ainsi que des données n'ayant pas de rapport avec la requête. Le système produit, en sortie, plusieurs ensembles d'images.

Ces ensembles sont au même nombre que celui des espaces de thèmes appris. Un ensemble d'espaces thématiques de granularités différentes est entraîné par une (LDA) sur un grand corpus de documents issus du Web. Un classifieur SVM, pour chacun des espaces de thèmes, est ensuite appris en utilisant des exemples positifs (un ensemble de photos pertinentes extraites du moteur de recherche Flickr) ainsi que des exemples négatifs (l'ensemble d'images considérées comme les moins pertinentes par le module *WEB*). Toutes les images annotées manuellement lors de la campagne MediaEval sont projetées dans chacun des espaces de thèmes puis sont traitées par le classifieur SVM associé à cet espace thématique.

Enfin, un vote à l'unanimité entre les classifieurs SVM permet de décider de la pertinence éventuelle de l'image. Cette dernière combinaison permet d'extraire un nouvel ensemble d'images considérées comme pertinentes.

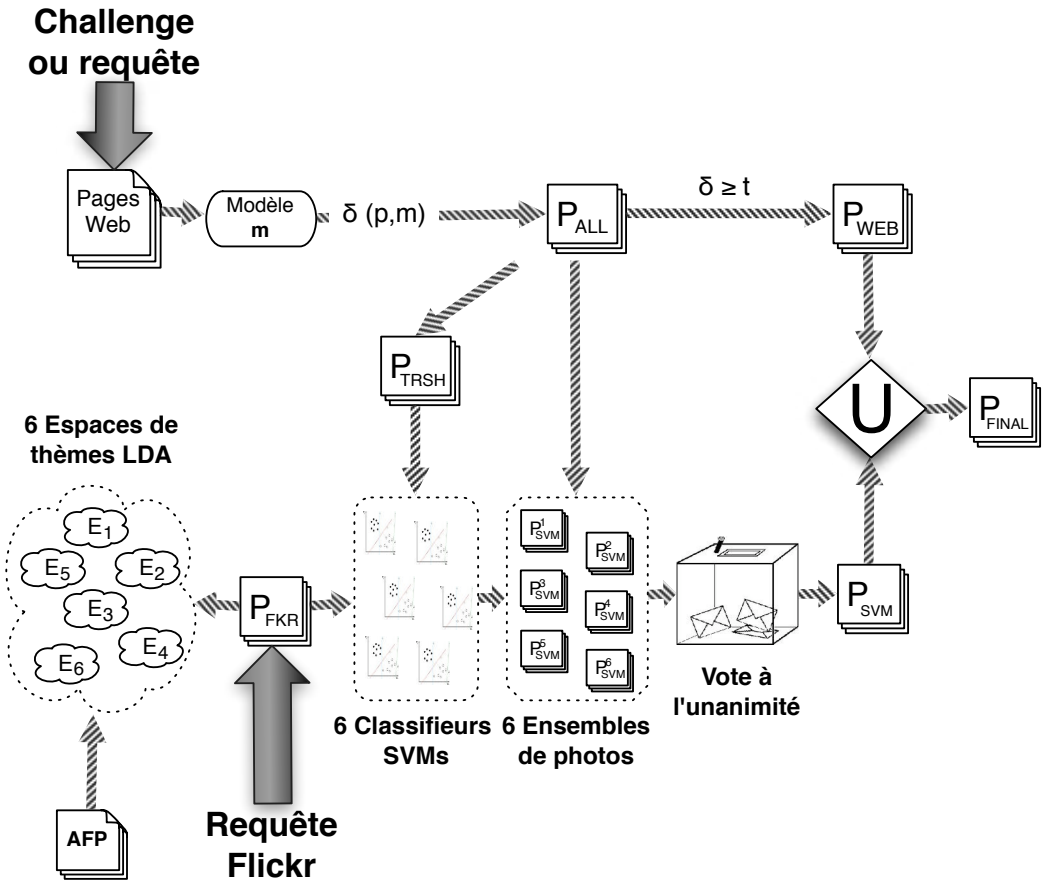


FIGURE 3.1 – *Architecture du système proposé pour l'extraction de photos répondant à une requête donnée.*

Finalement, un dernier processus réalise l'union entre les deux sous-ensembles des images pertinentes sélectionnées par le premier module *WEB* et l'ensemble de photos ayant récolté l'unanimité entre les différents classifieurs *SVM*. Ce dernier ensemble

de photos est considéré comme répondant au mieux à la requête. Le système de détection d'événements fondé sur une représentation multi-vues que nous proposons est présenté dans la figure 3.1.

Système fondé sur la fréquence des termes (*WEB*)

Ce processus en deux étapes extrait un sous-ensemble de photos pertinentes sachant une requête. La première étape consiste à récupérer un ensemble de pages Web répondant à cette requête. Le calcul de la fréquence des termes est réalisé pour tous les termes composant ces pages Web. Dans un second temps, chacune des images est évaluée en utilisant une mesure de similarité entre les tags composant l'image et les fréquences de termes du modèle. Ensuite, l'ensemble d'images est partitionné en deux sous-ensembles en fonction du score obtenu par la mesure de similarité : en deçà d'un seuil déterminé t, l'image appartient au sous-ensemble d'images considérées comme pertinentes. Dans le cas contraire, l'image est placée dans l'ensemble de photos non-pertinentes. Le processus complet est détaillé dans les sections qui suivent.

a) Estimation du modèle issu du Web

L'estimation du modèle m bâti sur la fréquence des mots requiert un grand corpus de documents D. Ce corpus est composé de pages Web répondant au mieux à la requête (cette requête est fournie par les organisateurs de la campagne MediaEval 2011). De ce corpus sont extraits un ensemble de termes et leurs fréquences associées. Uniquement le premier site Web [1] répondant à la requête [2] sur le moteur de recherche Google, est sélectionné. Toutes les pages Web associées à cette URL composent le corpus de documents utilisé pour l'estimation du modèle *WEB*. Une fois le corpus D collecté, la probabilité qu'un terme w apparaisse dans le corpus D est calculée comme suit :

$$P(w|m) = \frac{|w|_D}{N_D} \tag{3.1}$$

où $|w|_D$ est le nombre d'occurrences du mot w dans le corpus D et N_D le nombre total de termes contenus dans D. Cet ensemble de mots est utilisé pour déterminer le degré de similarité entre les tags d'une image, et le modèle m.

b) Classification des images

L'ensemble des photos P_{ALL} est partagé en un sous-ensemble de photos dites *pertinentes* P_{WEB} et un sous-ensemble de photos jugées comme *non-pertinentes* sachant la mesure de similarité δ entre une image p et le modèle m. Un ensemble de N photos les moins pertinentes (P_{TRSH}) est extrait depuis le sous-ensemble de photos non-pertinentes. La décision est prise en fonction du score δ comparativement au seuil t estimé sur les données de développement (challenge 1). Le calcul de δ fait appel à

1. *http ://www.paradiso.nl*
2. *may 2009 venue paradiso*

l'équation 3.1 pour estimer la similarité entre une photo p et le modèle *WEB* m sachant tous les termes w composant l'image p. La probabilité de w sachant la photo p est calculée de la même manière que la similarité δ :

$$\delta(p,m) = \sum_{w \in p} P(w|m)(1 + P(w|p)) \tag{3.2}$$

Le seuil t est optimisé en utilisant une requête et un corpus de développement. Cette requête vient de la campagne MediaEval 2011 (Papadopoulos et al., 2011a). Ce challenge traite d'événements footballistiques prenant place soit à Rome, soit à Barcelone (voir section 3.2.1.2 b). Une image ayant une similarité avec le modèle m supérieure à t est considérée proche du modèle et donc représentative de son contenu. Sachant cela, le modèle est composé de pages Web représentatives de la requête. Les photos issues de l'ensemble total de départ (P_{ALL}) ayant obtenu un score élevé de similarité δ avec le modèle m, sont donc considérées comme pertinentes. Deux sous-ensembles de photos sont donc obtenus : les images pertinentes P_{WEB} ainsi qu'un sous-ensemble de N photos considérées comme les moins proches du modèle m, et donc les moins représentatives de la requête, cet ensemble étant appelé P_{TRSH}.

Module multi-granulaires fondé sur une classification *SVM*

Dans le premier module *WEB*, l'ensemble de photos de départ P_{ALL} est divisé en un sous-ensemble de photos pertinentes et non-pertinentes. Dans cette section, une description est donnée du second module permettant de "repêcher" une partie des photos rejetées illégitimement par le premier module *WEB*.

La première étape consiste à définir un ensemble d'espaces de thèmes de granularités différentes. Un classifieur est ensuite entraîné pour chacun des espaces de thèmes. Ces classifieurs permettent de récupérer, dans l'ensemble des photos candidates P_{ALL}, celles étant les plus pertinentes. Ensuite, un vote est réalisé pour ne garder que les images ayant obtenu la majorité des votes de tous les classifieurs. Ce processus est décrit pas-à-pas dans les sections suivantes.

a) Représentation multi-granulaires s'appuyant sur les espaces de thèmes

Toutes les méthodes permettant de constituer un espace de thèmes requièrent un grand corpus de documents pour estimer de manière efficace les paramètres du modèle. Ici, le corpus d'apprentissage est constitué d'un ensemble de méta-données (ou *tags*) extrait des photos pertinentes, ainsi que d'articles de journaux ne reflétant pas le contenu de la requête. La requête en question est, dans un premier temps, envoyée au moteur de recherche du site Flickr via son API, pour récupérer un ensemble de 8 000 photos pertinentes, appelé P_{FKR}. Un second ensemble de données est composé d'articles issus de l'**A**gence **F**rançaise de **P**resse (AFP) écrits entre 2000 et 2006, dans les mêmes proportions que les images récupérées depuis le site Flickr (8 000). Finalement, un corpus de 16 000 documents est utilisé, après avoir été lemmatisé avec TreeTagger (Stein et Schmid, 1995), pour estimer un ensemble de 6 espaces de thèmes

LDA. Le nombre d'espaces thématiques ainsi que le nombre de thèmes les composant, sont déterminés pour fournir suffisamment de variété en termes de granularité (Blei et al., 2003; Rosen-Zvi et al., 2004). Ensuite, un ensemble de 6 modèles E_i est obtenu, contenant un nombre de dimension de l'espace de thèmes (10, 20, 30, 50, 100 et 200 thèmes dans nos expérimentations).

b) Machine à vecteurs de support (SVM)

Un SVM binaire (deux classes) est ici appris pour classifier en deux sous-ensembles (*pertinent* et *non-pertinent*) les photos projetées dans chacun des 6 espaces de thèmes LDA. Un ensemble de photos récupérées depuis Flickr P_{FKR} (classe $+1$), ainsi qu'un ensemble de N photos rejetées étant considérées comme les moins pertinentes sachant la mesure de similarité δ (P_{TRSH}) (classe -1), composent l'ensemble d'entraînement du classifieur SVM. Un ratio de $1 : 5$ (1 photo pertinente pour 5 non-pertinentes) est choisi pour cette tâche fortement déséquilibrée. Cette configuration obtient de meilleurs résultats qu'une configuration modérément déséquilibrée ($1 : 2$) ou équilibrée ($1 : 1$) (Kiritchenko et Matwin, 2001). Ainsi, un corpus de 8 000 images pertinentes (classe $+1$) et N =40 000 images moins pertinentes sachant le score de similarité δ (classe -1), est obtenu.

De plus amples détails concernant cette méthode de catégorisation sont disponibles dans la section 2.2.4.1.

c) Représentation des images dans un espace de thèmes LDA

Un SVM est appris pour chacun des 6 espaces de thèmes E_i de taille m_k thèmes. Chacun des documents d issu de P_{FKR} ou issu des photos les moins pertinentes (P_{TRSH}), est représenté par un vecteur V_d composé de m_k éléments $V_d[j]$ ($1 \leq j \leq n_z$). Ces éléments représentent la similarité du document d avec le thème z_j appartenant au modèle thématique E_i. L'équation suivante permet de calculer chaque composante du vecteur de représentation V dans un espace de thèmes E_i :

$$V_d[j] = \delta(d, z_j) \tag{3.3}$$

Chaque classifieur SVM_i est appliqué sur l'ensemble des images P_{ALL}. Un ensemble de photos pertinentes P^i_{SVM} est obtenu pour chaque espace de thèmes E_i.

d) Sélection d'un ensemble de photos pertinentes par vote à l'unanimité

Un sous-ensemble P_{SVM} est extrait depuis tous les sous-ensembles de photos associés à chacun des classifieurs SVM_i par l'intermédiaire d'un vote à l'unanimité entre tous les classifieurs. Ainsi, une photo appartenant à tous les sous-ensembles, est considérée comme pertinente. La diversité de représentation par des espaces de thèmes de tailles différentes, permet une représentation multi-granulaires pour décrire le contenu d'une photo *via* les méta-données qui lui sont associées. Par exemple, pour

la description d'un concert, un utilisateur peut utiliser les termes *groupe*, le nom du groupe, la compagnie de disque ou bien le style musical plus généralement (voir figure 3.2).

Union P_{WEB} et P_{SVM}

Deux sous-ensembles d'images (P_{WEB} et P_{SVM}) sont obtenus en utilisant les modules WEB et SVM. L'union de ces deux sous-ensembles est finalement réalisée pour obtenir un dernier ensemble nommé P_{Final} représentant au mieux la requête initial (challenge 2). Une image p est choisie selon la règle :

$$\begin{cases} p \in P_{Final} & \text{si} \quad p \in (P_{WEB} \cup P_{SVM}) \\ p \notin P_{Final} & \text{si} \quad p \notin (P_{WEB} \cup P_{SVM}) \end{cases}$$

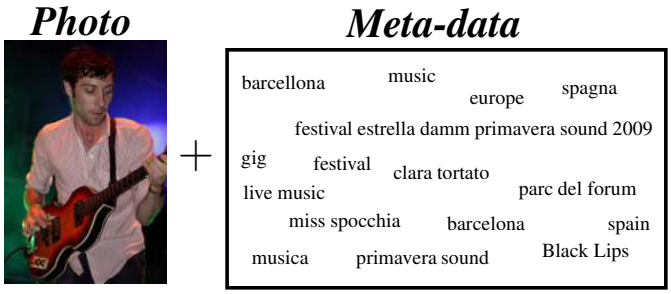


FIGURE 3.2 – *Exemple d'une photo issue du corpus MediaEval 2011 ainsi que les méta-données associées.*

3.2.1.2 Protocole expérimental

La méthode proposée est évaluée dans le cadre expérimental de la campagne d'évaluation MediaEval 2011. Les sections suivantes décrivent de manière plus détaillée cette campagne.

a) Campagne MediaEval 2011

En 2010, l'atelier VideoCLEF (Larson et al., 2010) devient MediaEval. Cette nouvelle campagne a pour but de permettre aux chercheurs de confronter leurs systèmes et leurs approches lors de "joutes" scientifiques autour du traitement de données multimédias (audio, image, vidéo, ...). Une tâche phare, apparue lors de la campagne d'évaluation MediaEval 2011, est la recherche d'événements sociaux dans un ensemble de données bruitées issues de plateformes de partage (Papadopoulos et al., 2011b). Notre système multi-vues a été évalué sur cette tâche.

b) Corpus de MediaEval 2011 concernant la recherche d'événements sociaux

Un ensemble de 73 645 photos (P_{ALL}) est collecté depuis l'API de la plateforme de partage d'images Flickr [3]. Cet ensemble constitue la totalité des photos prises en mai 2009, contenant des coordonnées de géolocalisation disponibles sur la plateforme pour cinq villes : Amsterdam, Barcelone, Londres, Paris et Rome. À cet ensemble de photos, sont ajoutées d'autres images ne possédant pas de coordonnées de géolocalisation pour les mêmes villes et à la même période en utilisant le système décrit dans (Troncy et al., 2010).

La figure 3.3 montre quelques exemples de photos contenues dans le corpus MediaEval 2011 pour la tâche de détection d'événements sociaux. Dans le corpus de photos ainsi obtenu, les coordonnées de géolocalisation sont retirées pour 80 % des photos sélectionnées aléatoirement. Ceci est réalisé pour simuler artificiellement le manque d'information géographique observé dans la plupart des documents provenant du Web [4].

FIGURE 3.3 – *Exemple de photos appartenant à l'ensemble de photos du corpus MediaEval 2011 et concernant (a) des événements à Paradiso à Amsterdam, (b) des événements musicaux à Parc del Forum à Barcelone, ou (c) des événements footballistiques dans les villes de Rome ou de Barcelone.*

c) Tâche de détection d'événements sociaux (SED)

Les participants à cette tâche ont pour but de "découvrir" un ensemble d'évé-

3. **https://www.flickr.com**
4. Le corpus, ainsi qu'un descriptif de la tâche SED et des autres tâches de la campagne MediaEval 2011, sont disponibles ici : **http://www.multimediaeval.org/mediaeval2011/SED2011/**

nements, puis d'associer à chacun des événements, un ensemble de documents multimédias (images dans notre cas). L'utilisateur a souvent tendance à poster des documents saisis ou liés à un événement précis. De plus, un utilisateur d'un moteur de recherche de documents multi- ou mono-médias recherchera un document ou un ensemble de documents en fonction d'une requête liée à un événement précis. Pour ces raisons, la nécessité de créer des relations bi-directionnelles entre les documents et les événements qu'elles décrivent, est primordiale si l'objectif du partage de telles quantités de données est de les trouver "naturellement".

La tâche de détection d'événements sociaux (SED) est composée de deux challenges ainsi que d'un ensemble de test commun à ces deux challenges. Chacun des éléments composant les corpus d'apprentissage et de validation, est composé d'images et de méta-données leur étant associées (dont les dates de saisies, étiquettes, et coordonnées de géolocalisation pour certaines d'entre elles). La tâche de SED est composée de deux challenges correspondant à deux requêtes bien distinctes :

Challenge 1

Trouver tous les événements footballistiques ayant pris place dans les villes de Barcelone (Espagne) ou Rome (Italie) dans le corpus de validation. Pour chacun des événements trouvés, fournir toutes les photos associées à celui-ci.

Challenge 2

Trouver tous les événements ayant eu lieu en mai 2009 au lieu nommé Paradiso (Amsterdam, Pays-Bas) ou au Parc del Forum (Barcelone, Espagne). Pour chacun des événements, fournir toutes les photos associées à celui-ci.

L'objectif de la tâche 1, est de récupérer les 1 640 images représentant le challenge et non de trouver l'ensemble des événements. Le système proposé dans cette étude utilise uniquement les méta-données liées à la photo. Ceci entraîne des erreurs sur les annotations faites par l'utilisateur. La figure 3.4 montre une image issue du corpus MediaEval 2011. Si l'on ne se réfère uniquement qu'aux méta-données, tout porte à penser que cette photo traite du challenge 1 concernant un événement footballistique. Ceci n'est évidement pas le cas, puisque la photo qui est associée à cet ensemble de méta-données montre le plafond de la chapelle sixtine à Rome. Parmi les méta-données fournies par l'utilisateur, seul le terme Rome semble en relation avec l'image.

La requête utilisée pour évaluer la pertinence d'une représentation multiple d'un document bruité issu du Web, est le challenge 2. Les requêtes *"paradiso amsterdam"* ainsi que *"parc del forum barcelona"* sont envoyées à la plateforme Flickr *via* son API pour collecter un ensemble de 8 000 photos pertinentes P_{FKR}. Ces photos constituent le corpus permettant, en partie, l'apprentissage des espaces de thèmes LDA ainsi que la constitution du corpus de documents positifs (classe $+1$) lors de l'apprentissage des classifieurs SVM.

Meta-données **Photo**

</description>
<tags>
<tag>**Champions**</tag>
<tag>**League**</tag>
<tag>**Rome**</tag>
<tag>**Stadium**</tag>
<tag>**Football**</tag>
<tag>**Manchester**</tag>
<tag>**Barcelona**</tag>
<tag>**Champions League**</tag>
<tag>**Manchester United**</tag>
</tags>
<location latitude="41.90311"
longitude="12.495759"></location>
</photo>
<photo id="3517300861" photo_url="http://
farm4.static.flickr.com/
3570/3517300861_5e4f05f74f.jpg"
username="veensantmo"
dateTaken="2009-05-07 11:28:40.0">
<title>20090507-112840</title>
<description>

FIGURE 3.4 – *Exemple d'une photo appartenant au corpus MediaEval 2011.*

3.2.1.3 Résultats et discussions

Le premier module permet d'identifier respectivement 1 612 (P_{WEB}) et 72 033 images correspondant aux photos pertinentes et non-pertinentes. Ceci est réalisé en utilisant un seuil t de $0,133$ estimé sur les données de développement. La méthode s'appuie uniquement sur une collection de pages Web, sans l'utilisation d'un corpus annoté pour l'apprentissage d'un modèle fondé sur les fréquences de mots. Lorsque peu de données sont disponibles pour la phase d'apprentissage, les informations issues du Web sont un bon point de départ pour la constitution d'un corpus de données plus hétérogènes. Celui-ci permet d'atteindre un niveau de performance satisfaisant, tout en ne disposant que de peu de données au départ. En effet, avec cette approche, plus de 68 % des images représentant une requête donnée peuvent être correctement récupérées. Il est également à noter que la mesure permettant l'évaluation (voir équation 3.2) permet d'associer plus de poids aux termes apparaissant simultanément dans les "contenus" lexicaux de l'image, *via* ses méta-données, et dans les pages Web décrivant la requête. Cependant, un inconvénient de cette métrique de similarité entre une photo et le modèle Web fondé sur les fréquences de mots, est l'absence d'une pondération prenant en compte l'importance du terme au sein des documents composant le corpus d'apprentissage (comme l'est l'IDF (Inverse Document Frequency) (Salton, 1989a)).

Dans le second module, un classifieur SVM est appris pour chacun des 6 espaces de thèmes avec l'ensemble annoté de photos fourni par les organisateurs de la campagne SED de MediaEval 2011. L'ensemble P_{TRSH} utilisé lors de la phase d'apprentissage des SVM, est constitué de 40 000 photos les moins pertinentes classifiées avec le premier

module *WEB*. Le tableau 3.1 présente les résultats obtenus durant la tâche d'extraction pour chacun des espaces de thèmes avec un classifieur SVM.

#thèmes	#trouvés	#justes	Précision	Rappel	F-mesure
10	6 822	1 255	18,4	76,5	30,0
20	6 811	1 377	20,2	84,0	32,6
30	6 076	1 264	20,8	77,1	32,8
50	8 006	1 315	16,4	80,2	27,2
100	7 744	1 127	14,5	68,7	24,0
200	7 304	1 417	19,4	86,4	31,6
Vote	**395**	**218**	**55,2**	**13,3**	**21,4**

TABLE 3.1 – *Résultats obtenus avec le module SVM pour chacun des 6 espaces de thèmes*

Les résultats montrent qu'une classification utilisant un seul espace de thèmes, obtient un rappel élevé (de 76,5 % à 86,4 %), mais avec une précision relativement faible (de 20,8 % à 14,5 %). Comme attendu, le processus de vote à l'unanimité améliore notablement la précision. La combinaison des deux modules (union), correspondant à la maximisation successive du rappel et de la précision, obtient des résultats encourageants. Le tableau 3.2 rapporte les performances des deux modules (*WEB* et *SVM*) pris séparément, ainsi que leur union en termes de précision, rappel et F-mesure. En particulier, il est à noter que 372 images sont rejetées (jugées comme non-pertinentes) par l'union des deux systèmes.

Méthode	#trouvés	#justes	Précision	Rappel	F-mesure
WEB (1)	1 612	1 108	67,6	68,8	68,2
SVM (2)	395	218	55,2	13,3	21,4
$1 \cup 2$	**1 900**	**1 268**	**77,3**	**66,7**	**71,6**
Meilleur système	1 737	1 164	67,0	71,0	69,9

TABLE 3.2 – *Performances des deux modules WEB et SVM, leur union, ainsi que le système ayant obtenu les meilleurs résultats lors de la campagne SED de MediaEval 2011.*

La contribution de cette représentation dans des espaces de thèmes multiples n'est pas négligeable. Comme il est indiqué dans le tableau 3.2, le score de F-mesure est augmenté de 3,4 points (de 68 % à plus de 71 %). Le nombre de photos pertinentes trouvées augmente lui de 10 points quand sont comparés les résultats obtenus avec le simple système *WEB* ou le meilleur système proposé lors de la campagne MediaEval 2011 avec l'union des modules *WEB* et *SVM* (score de F-mesure de 68,95 %) (Liu et al., 2011).

Finalement, l'utilisation de cette multiple représentation et de classifieurs SVM permet de récupérer 160 photos pertinentes supplémentaires rejetées par le système bâti sur la fréquence de mots *WEB*. En effet, 1 268 photos sont correctement trouvées lorsque l'on procède à l'union des deux modules alors que seulement 1 108 sont considérées comme pertinentes par le module *WEB*. En fait, le module *SVM* est composé de 73 % de photos n'appartenant pas au premier sous-ensemble de photos pertinentes obtenu

par le module *WEB*. Ceci valide l'idée basique que cette méthode s'appuyant sur une représentation multi-vues d'un document bruité, permet de récupérer des images incorrectement rejetées par l'approche faiblement supervisée bâtie sur les fréquence de mots dans un corpus de pages Web. Il est également à observer, dans le tableau 3.3, que les images qui n'ont pas été trouvées par le système fondé sur l'union des deux modules, ont une structure de méta-données qui dépasse le simple bruit.

Images trouvées	Images rejetées
audience primavera barcelona	junk food 難食到呢
2009 sound primavera barcelona	cimg0367
img7584 2009 sound primavera	entrance

TABLE 3.3 – *Méta-données associées à des photos trouvées et rejetées à tort par l'union des deux modules WEB et SVM.*

Globalement, les résultats justifient l'approche proposée consistant à pallier la faiblesse de la structure des méta-données, en utilisant une représentation multiple abstraite de haut-niveau (pas uniquement lexicale comme *WEB*). Notre approche permet d'améliorer significativement les résultats obtenus par rapport à une approche plus simple basée sur les fréquences de mots.

3.2.1.4 Conclusions sur la représentation multi-granulaires de documents bruités issus du Web

Dans cette étude, une méthode robuste d'extraction de photos depuis une requête utilisant une représentation multiple est proposée. Cette méthode fournit une nouvelle alternative à la caractérisation d'une photo, non seulement en utilisant les méta-données lui étant associées, mais avec une représentation s'appuyant sur les thèmes proches dans des espaces thématiques multiples. Ces espaces de thèmes sont appris avec une allocation latente de Dirichlet (LDA) utilisant un corpus de photos pertinentes et des articles de journaux choisis aléatoirement. Différents sous-ensembles d'images provenant de divers espaces de thèmes sont combinés pour extraire un ensemble d'images représentant une requête. Les expérimentations montrent la contribution de cette représentation de haut-niveau proposée, avec une meilleure F-mesure de 71 % qui améliore les résultats obtenus avec l'utilisation du système basique seulement. Ceci montre que la représentation multiple d'un document bruité dans des espaces de thèmes de tailles différentes, permet une indexation robuste d'images comparativement à une représentation lexicale simple. Ces résultats soulignent également que la représentation dans un espace de thèmes permet un filtrage du contenu textuel du document (méta-données).

Le système proposé dans cette étude est uniquement évalué pour une requête donnée afin de récupérer des photos dans un ensemble de photos fourni lors de la campagne MediaEval 2011. Pour permettre une généralisation d'une telle représentation à d'autres requêtes et à un grand corpus de photos, cette approche doit être évaluée avec différents types de données (texte, courriel, ...) contenant plus de variétés.

Le dernier point de la discussion ouvre des perspectives pour améliorer ces résultats en utilisant d'autres caractéristiques en complément des méta-données seules. Ainsi, les informations multimédias contenues dans une image ne sont pas encore exploitées : par exemple, le contenu de la photo (traitement de l'image) est une voie à explorer pour trouver des photos pertinentes non détectées, et raffiner l'extraction en séparant les images selon l'événement qu'elles décrivent.

Cette représentation multi-vues a montré son efficacité dans une tâche de classification de documents bruités issus du Web. De la même manière que dans le chapitre 2, la méthode proposée fondée sur une représentation multiple d'un même document, est évaluée sur des documents bruités issus de transcriptions automatiques de documents audios. La section 3.2.2 présente une étude d'une telle représentation multiple sur une tâche de reconnaissance du thème principal d'une conversation entre un agent et un utilisateur de la RATP.

3.2.2 Représentation multi-thèmes de documents parlés transcrits automatiquement pour une catégorisation robuste

Une représentation multiple d'un document textuel bruité provenant du Web permet, en plus d'une abstraction forte du contenu du document par une représentation en ensemble de thèmes proches, de prendre en compte les différents niveaux de granularité dudit document. La section 2.2 du chapitre 2 a présenté une étude montrant l'intérêt d'une représentation thématique d'un document parlé lors d'une tâche de catégorisation. De la même manière, l'étude présentée dans cette section évalue l'apport d'une représentation multiple d'un document parlé dans différents espaces thématiques de granularités différentes. Cette étude se place dans le même contexte décrit précédemment, du corpus DECODA et de la catégorisation de dialogues.

L'identification du thème principal est réalisée dans le cadre de conversations réelles entre un agent et un utilisateur de la RATP. Dans ce contexte, malgré la difficulté prévisible liée à la transcription de documents enregistrés dans des conditions souvent difficiles (rue, métro, ...), une précision élevée est observée durant la tâche de recherche du thème principal (section 2.2 du chapitre 2). Ces résultats sont obtenus avec un classifieur probabiliste utilisant des caractéristiques obtenues avec un espace de thèmes cachés LDA commun à l'agent et à l'utilisateur.

Cette partie prolonge l'étude du chapitre 2 en appliquant, au même problème, notre représentation multi-vues.

La figure 3.5 présente un dialogue entre un utilisateur et un agent de la RATP, et les deux parties composant le dialogue (utilisateur et agent). Des espaces de thèmes spécifiques sont considérés respectivement pour l'agent et pour l'utilisateur ainsi que la combinaison des deux représentations. De plus, comme différents environnements bruités peuvent conduire à une variation très grande du taux d'erreur-mot (TEM) pour chacune des conversations, nous considérons la possibilité de prendre en compte le TEM pour chacune des conversations contenues dans le corpus d'apprentissage lors de l'apprentissage du classifieur.

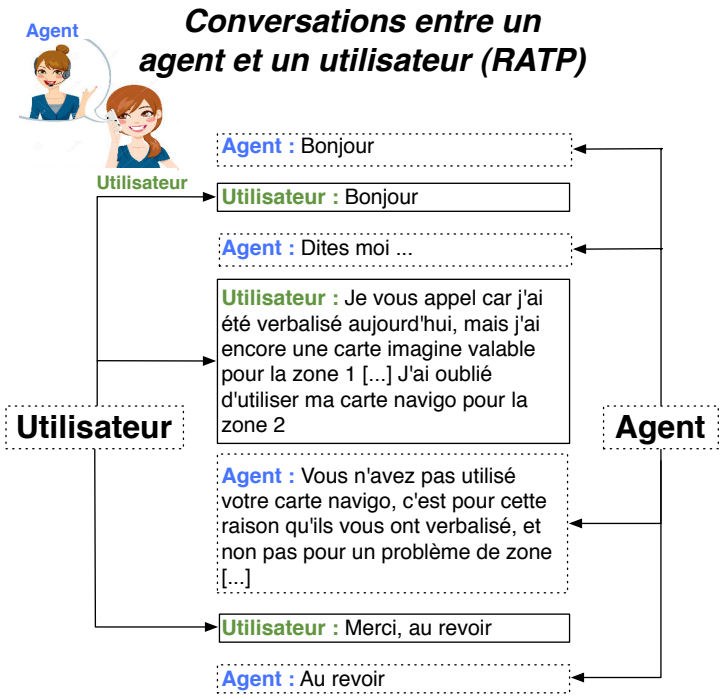


FIGURE 3.5 – *Exemple d'un dialogue entre un utilisateur et un agent de la RATP.*

Les résultats en termes de recherche du thème principal sont obtenus avec trois représentations thématiques différentes ainsi que différentes granularités (nombre de thèmes composant l'espace de thèmes) en utilisant un classifieur gaussien s'appuyant sur une règle bayésienne. Cette étude exploite partiellement ou totalement les résultats obtenus avec des espaces de thèmes différents. Ayant des processus de classification séparés pour chacune des représentations des dialogues dans un espace de thèmes, il est alors possible de comparer les résultats obtenus pour diverses granularités et d'utiliser le consensus partiel ou total entre tous les classifieurs comme un indicateur sur le thème principal à associer à un dialogue.

Une amélioration importante, comparativement aux résultats observés dans l'étude de l'apport d'une représentation thématique d'un dialogue (voir section 2.2 du chapitre 2), est obtenue en utilisant une stratégie simple suggérée par les résultats obtenus avec l'ensemble de développement. Elle consiste à sélectionner le thème ayant obtenu le consensus maximum sur les trois représentations thématiques, ou le thème trouvé, en utilisant l'espace de thèmes propre à l'agent dans le cas où les trois classifieurs sont en total désaccord.

Pour la faible proportion des conversations classifiées d'une manière ambiguë, une simple mesure de confiance fondée sur le consensus entre les différents classifieurs est utilisée pour signaler si l'intervention d'un expert humain est nécessaire pour obtenir une classification fiable avec un effort moindre.

La section 3.2.2.1 qui suit présente la représentation multi-vues d'un même dia-

logue, ainsi que les différentes méthodes de classification. La section 3.2.2.2 présente le protocole expérimental. Les résultats, en termes de bonne catégorisation, sont donnés dans la section 3.2.2.3 avant de conclure sur l'apport d'une représentation multiple d'une transcription fortement bruitée dans la section 3.2.2.4.

3.2.2.1 Approche proposée pour une représentation multi-vues d'une transcription fortement bruitée

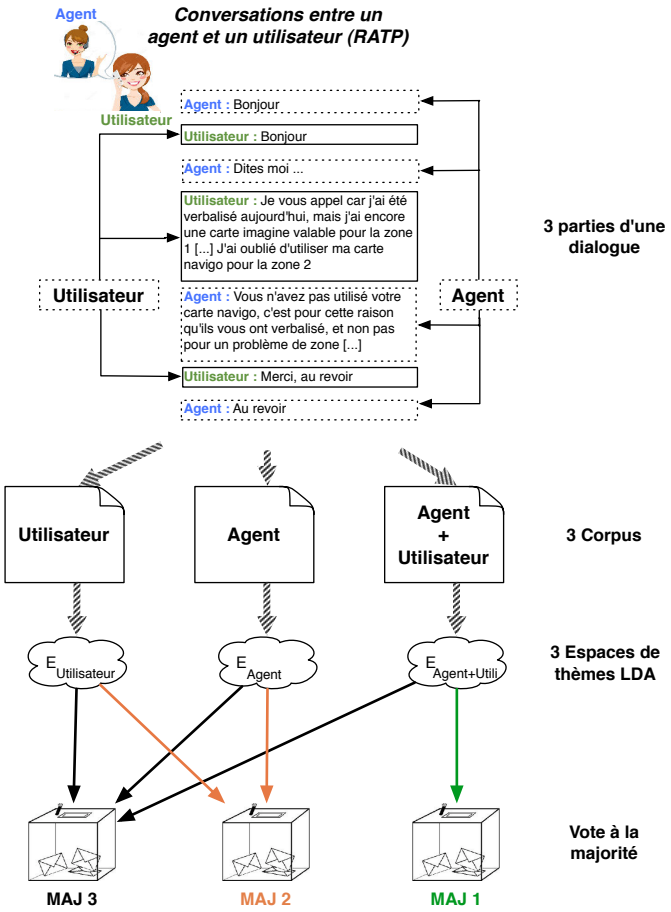


FIGURE 3.6 – *Approche proposée pour une représentation multiple et une catégorisation robuste d'un dialogue issu du corpus DECODA.*

La figure 3.6 montre les différents processus du système proposé :

- Découpage du corpus de dialogues en 3 sous-corpus composés de la partie provenant de l'utilisateur, de l'agent et l'association des deux (corpus original).
- Création d'un espace de thèmes spécifique à chacun des corpus.
- Projection de chacun des dialogues dans les trois espaces thématiques précédemment élaborés.

- Classification de la représentation de chacun des dialogues dans chacun des trois espaces de thèmes.
- Évaluation si les trois classifieurs soient d'accord (MAJ3), si uniquement deux classifieurs proposent le même thème principal (MAJ2), ou si les trois classifieurs suggèrent un thème différent (MAJ1). Dans ce dernier cas, le thème choisi est celui proposé par le classifieur issu du dialogue entier.

Une fois la première étape de constitution des trois corpus effectuée, le système élabore un espace de thèmes en utilisant chacun des corpus à l'aide d'une LDA (voir annexe D). Chacun des dialogues est ensuite projeté dans chacun des espaces de thèmes. Ainsi, une représentation vectorielle du dialogue dans l'espace thématique est obtenue. Cette représentation, utilisant un vocabulaire discriminant réduit, est détaillée dans la section 2.2.3.2 du chapitre 2. Cette représentation vectorielle du dialogue est fournie en entrée d'un classifieur gaussien introduit dans la section 2.2.4.2 du chapitre 2 également.

3.2.2.2 Protocole expérimental

Le corpus de dialogue entre un utilisateur et un agent de la RATP du projet DECODA (Bechet et al., 2012) est utilisé pour les expérimentations décrites dans cette section. La section 2.2.5 présente le corpus en détail.

Le nombre de tours de parole dans une conversation et le nombre de termes dans chacun de ces tours de parole varient fortement. La majorité des conversations contient plus de 10 tours de parole. Les tours de parole des utilisateurs tendent à être plus longs ($>$ 20 mots) que ceux des agents, et sont plus enclin à contenir des termes non-pertinents pour cette tâche de catégorisation.

Le système de reconnaissance automatique de la parole, nécessaire pour extraire les transcriptions automatiques depuis les documents audios, est le système SPEERAL (Linarès et al., 2007) décrit dans la section 2.2.5 ainsi que dans l'annexe E.

Les expérimentations sont conduites en utilisant le corpus d'apprentissage représenté par les transcriptions manuelles uniquement (TMAN), les transcriptions automatiques uniquement (TRAP), ainsi que les transcriptions automatiques tenant compte du taux d'erreur-mot (TRAP+TEM).

En effet, nous évaluerons l'impact du TEM durant la tâche de classification. $\overline{x_k}$, qui permet de définir la règle de décision lors du calcul de la distance de Mahalanobis décrite dans l'équation 2.4 de la section 2.2.4.2, est à présent calculé sur le corpus d'apprentissage de la façon suivante TRAP :

$$\overline{x_k^{wer}} = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{x_k^i}{wer_i} , \tag{3.4}$$

avec wer_i le taux d'erreur-mot (TEM) pour le i^{th} dialogue.

Les conditions indiquées par les abréviations entre parenthèses sont les ensembles de développement (Dev) et de validation (Test). La tâche de classification est réalisée

dans les conditions d'apprentissage et de validation décrites auparavant en utilisant un vecteur de caractéristiques issu des espaces de thèmes provenant respectivement de l'agent seul (AGENT), de l'utilisateur seul (UTILISATEUR) et de la combinaison des deux (AG+UT).

Pour permettre une bonne comparaison, la tâche de recherche da la catégorie principal dans les dialogues du corpus DECODA fait appel à deux méthodes de classification (SVM et gaussien ou Mahalanobis) et un vecteur de caractéristiques contenant la probabilité de chacun des mots discriminants. Il est également présenté les résultats obtenus avec une représentation utilisant des caractéristiques fondées sur le TF-IDF-Gini (voir section 2.2.3.1) avec un classifieur SVM (voir section 2.2.4.1) pour comparaison.

Comme l'agent annote une conversation comme appartenant au thème qu'il juge comme le plus important, si cette conversation traite de plusieurs thèmes, le corpus d'annotation utilisé lors des expérimentations s'appuie sur les annotations de l'agent. Ces annotations peuvent cependant faire l'objet de corrections mineures uniquement lorsque des erreurs incontestables de l'agent, dues par exemple au stress, sont observées.

Des conditions expérimentales identiques à celles présentées dans le chapitre 2.2.5 sont utilisées pour mener à bien nos expériences. Un sous-ensemble de 800 mots discriminants sont extraits pour composer l'ensemble des caractéristiques des vecteurs de représentation des dialogues, appelés TF-IDF-Gini, combinés avec un classifieur SVM. Cette approche constitue notre système de base (ou *baseline*) qui sera comparé au système multi-vues proposé dans cette étude. Pour chacune des conditions et types de caractéristique, un ensemble de 19 espaces de thèmes cachés LDA de tailles différentes ($\{5, 6, 7, \ldots, 10, 20, \ldots, 100, 150, \ldots, 300\}$) est construit en utilisant le corpus d'apprentissage (Train).

3.2.2.3 Résultats obtenus lors de la catégorisation de représentations multi-vues de dialogues issus de transcriptions automatiques

Les résultats obtenus lors des expérimentations utilisant des caractéristiques issues d'espaces de thèmes et le classifieur gaussien sont présentés dans le tableau 3.4.

Chacune des lignes correspond à une configuration différente des données d'apprentissage (Train) et de validation (Dev/Test). Chacune des colonnes correspond à un type de locuteur (AGENT,UTILISATEUR ou AG+UT) et un type d'ensemble (Dev/Test) sur lequel l'évaluation est réalisée.

Les nombres reportés dans les colonnes étiquetées avec $|t|$ indiquent la taille de l'espace de thèmes cachés LDA correspondant au milieu de l'intervalle de $|t|$ dans lequel la performance de catégorisation varie le moins quand le corpus de validation est l'ensemble de développement. Une stratégie de consensus très simple est ensuite appliquée. Celle-ci consiste en la sélection de la catégorie récoltant le score le plus haut lors de la catégorisation pour au moins deux espaces de thèmes cachés. Ce score permet de sélectionner l'hypothèse générée en utilisant uniquement les caractéristiques issues de

l'espace de thèmes AG+UT quand il n'y a pas de consensus entre les trois classifieurs. L'intervalle de confiance pour l'ensemble de validation (Test) est de $\pm$ 3,69 %.

Comme attendu, les résultats montrent que les meilleures performances, utilisant des transcriptions TRAP pour l'ensemble de développement (Dev), sont obtenues avec des dialogues TRAP pour la phase d'apprentissage. Les résultats montrent également qu'il n'y a aucun avantage à retirer de l'utilisation d'une information au niveau du TEM lors de la phase d'apprentissage du classifieur gaussien, ou du moins, dans la forme simple utilisée ici. En effet, le tableau 3.4 montre que la précision de catégorisation n'est pas meilleure lorsque le TEM est considéré dans l'estimation de $\overline{x_k}$ (TRAP→TRAP avec TEM, voir équation 2.4) que lorsque l'on utilise le centroïde $\overline{x_k}$ (TRAP→TRAP, voir équation 3.4).

De plus, la précision observée pour le cas de l'AGENT seul est plus élevée que celles observées pour l'UTILISATEUR ou l'association des deux. Ceci peut-être dû à la différence d'environnement et de lexique utilisé entre l'agent et l'utilisateur. L'influence de ces différences est en quelque sorte atténuée par le fait que les caractéristiques sont extraites depuis les représentations abstraites que sont les thèmes composant l'espace LDA spécifique à chacun des types de locuteurs.

Une analyse du corpus d'apprentissage montre que les agents tendent à utiliser des expressions similaires pour expliquer certains points à l'utilisateur. Ceci est dû à l'entraînement que les agents suivent pour répondre au plus près au protocole de communication avec l'utilisateur défini par la RATP. Cependant, les utilisateurs répètent souvent les détails concernant leurs problèmes en fournissant des informations complémentaires.

Données	*UTILISATEUR*			*AGENT*			*AG+UT*		
Train/Test	$\|t\|$	**Dev**	**Test**	$\|t\|$	**Dev**	**Test**	$\|t\|$	**Dev**	**Test**
TMAN/TMAN	*80*	89,7	**84,7**	*80*	91,0	**85,3**	*80*	92,5	**86,8**
TRAP/TRAP	*80*	85,4	**79,4**	*70*	87,4	**80,4**	*80*	84,8	**82,5**
TRAP+TEM/TRAP	*80*	76,0	**72,3**	*60*	78,1	**73,1**	*100*	75,6	**75,2**
TMAN/TRAP	*100*	82,4	**74,4**	*100*	85,4	**77,6**	*80*	84,8	**78,3**

TABLE 3.4 – *Comparaison de performance, en termes de proportion de catégories trouvées, utilisant différents espaces de thèmes cachés pour différentes situations d'apprentissage et différents types de locuteurs.*

Le tableau 3.5 montre les résultats obtenus en utilisant un classifieur gaussien comparé à l'utilisation des même caractéristiques avec un classifieur SVM. Les expérimentations sont également conduites en utilisant des caractéristiques provenant de la liste de termes discriminants TF-IDF-Gini accompagnées d'un classifieur SVM.

La stratégie de recherche da la catégorie principal dans un dialogue MAJ1, consiste en la sélection du thème ayant recueilli le consensus maximum pour les trois espaces de thèmes associés aux trois cas : AGENT, UTILISATEUR et AG+UT. Dans le cas où les trois classifieurs associés aux trois espaces de thèmes livrent une décision différente, le thème retenu est celui du classifieur associé à l'AG+UT.

Les résultats obtenus avec l'approche proposée sont supérieurs à ceux observés dans (Morchid et al., 2014b,a). De plus, les améliorations de cette méthode s'appuyant sur des espaces de représentation multiples sont à présent statistiquement significatifs.

Données	TF-IDF-GINI	SVM	*AG+UT*	*MAJ1*
Train/Test	**Test**	**Test**	**Test**	**Test**
TMAN/TMAN	74,1	85,5	86,8	**87,2**
TRAP/TRAP	64,5	80,4	82,5	**84,1**
TMAN/TRAP	58,4	72,0	78,3	**78,5**

TABLE 3.5 – *Comparaison entre les précisions obtenues avec différents classifieurs utilisant des caractéristiques issues d'espaces de thèmes LDA.*

Une stratégie fondée sur le consensus des classifieurs dans des conditions différentes est utilisée pour composer les ensembles de validation suivants :

- MAJ3 est l'ensemble de test composé des conversations pour lesquelles la catégorie associée à la conversation est identique pour les trois classifieurs : AGENT, UTILISATEUR et AG+UT,
- MAJ2 est l'ensemble de test contenant toutes les conversations pour lesquelles au moins deux des trois classifieurs prédisent une catégorie identique,
- Dans le cas où les trois classifieurs choisissent une catégorie différente, la catégorie conservée est celle de AG+UT et est ajouté au deux sous-ensembles pour former l'ensemble MAJ1.

Le tableau 3.6 montre, pour chacun des types de consensus (MAJ3, MAJ2 et MAJ1), la précision du système de catégorisation de dialogues ainsi que la mesure de la couverture correspondant à la proportion de conversations considérées dans l'ensemble. Les résultats montrent que la stratégie s'appuyant sur le consensus fournit une bonne mesure de confiance bien qu'aucune stratégie spécifique ne soit employée lors de la phase d'apprentissage.

Données	*MAJ3*		*MAJ2*		*MAJ1*	
Train/Test	**Précision**	**Couverture**	**Précision**	**Couverture**	**Précision**	**Couverture**
TRAP/TRAP	95,1	*57,6*	86,3	*92,2*	**84,1**	*100*

TABLE 3.6 – *Précision lors de la catégorisation des dialogues et couverture pour chacune des stratégies dans les conditions d'apprentissage TRAP et l'ensemble de Test provenant de TRAP également (TRAP→TRAP) (espace de thèmes de taille 80) en %.*

3.2.2.4 Conclusions sur l'apport d'une représentation dans de multiples espaces de thèmes pour la catégorisation de transcriptions fortement imparfaites

Les résultats reportés dans le tableau 3.5 montrent que les caractéristiques issues d'espaces thématiques associées à un type de locuteur précis dans une conversation permettent de capturer le contenu sémantique utile avec des performances significativement supérieures à celles observées avec des caractéristiques indépendantes issues du TF-IDF-Gini. Nous observons également des gains significatifs par rapport à la technique fondée sur un unique espace de thèmes.

Il est aussi à remarquer que les résultats fournis dans le tableau 3.6 montrent que des mesures de confiance peuvent être estimées avec une stratégie simple, s'appuyant sur le vote, qui ne requiert aucune estimation des paramètres. L'utilisation d'une telle mesure pourrait permettre d'extraire un sous-corpus relativement grand (57 % dans notre expérience) pour lequel la précision est élevée (95 %), en dépit d'un TEM élevé. Une telle précision avec une aussi bonne couverture constitue un avantage concret assez clair en comparaison à d'autres études utilisant un corpus de documents similaires (Maza et al., 2011). De même, il devient possible d'estimer les proportions de problèmes rencontrés par les utilisateurs à un instant donné ou dans une situation de transport spécifique. Si ces proportions sont estimées dans une étude suffisamment étendue, les problèmes et les requêtes de l'utilisateur peuvent être surveillées pour prendre les décisions les plus pertinentes permettant d'améliorer le service fourni aux utilisateurs.

3.3 Conclusions générales sur la représentation multiple de documents bruités dans des espaces de thèmes

Le chapitre 2 a montré l'apport d'une représentation thématique pour différentes tâches de recherche d'information sur des documents bruités issus du Web ou de transcriptions imparfaites.

Dans ce chapitre, nous présentons un système de catégorisation de documents bruités s'appuyant sur l'extraction de vecteurs de représentation issus de vues multiples dans des espaces de thèmes LDA. Cette représentation permet au système de catégorisation de tolérer des erreurs de diverses natures, par exemple liées à l'annotation par des humains (voir section 3.2.1), ou des transcriptions très imparfaites (voir section 3.2.2).

Cette représentation multiple d'un document souffre néanmoins de plusieurs défauts :

- chaque projection d'un document ajoute une variabilité liée à l'espace de représentation (différence de granularité, distribution de mots dans les classes LDA différente, ...),
- la méthode oblige l'estimation d'un grand nombre grand de modèles LDA,
- la décision par vote est assez sommaire, il est probable qu'une stratégie de fusion plus sophistiquée permettrait de tirer un meilleur parti des différentes vues,
- le choix *a priori* des hyper-paramètres des modèles LDA n'est pas résolu (notamment des paramètres α et β).

Le problème de la variabilité additionnelle générée par l'approche multi-vues peut être traité par des méthodes d'analyse factorielle, qui sont des méthodes classiques d'analyse de données qui ont été récemment appliquées à des problèmes de traitement de la parole.

Dans le chapitre suivant, nous allons voir de quelles manières ces méthodes issues du traitement de la parole peuvent être appliquées à notre problème de variabilité intervues et de configuration des vues dans une approche multi-vues.

Chapitre 4

L'analyse factorielle pour une catégorisation robuste d'une représentation multiple compactée d'un document bruité

Sommaire

Résumé

Ce chapitre propose de poursuivre l'étude de l'apport d'une représentation multiple d'un document bruité dans des espaces de thèmes LDA. Ceux-ci souffrent de faiblesses inhérentes à la projection multiple dans des espaces de thèmes différents ainsi que de la difficulté d'une bonne estimation des paramètres du modèle LDA. Pour contourner la difficulté liée au choix des valeurs des paramètres du modèle, de nombreux espaces sont appris, mais ceux-ci introduisent une variabilité inutile liée à la multiplication des projections. Pour pallier cette faiblesse, les méthodes issues de l'analyse factorielle sont utilisées. En effet, les chercheurs du domaine de la vérification du locuteur ont déjà fait face à cette problématique liée à la variabilité nuisible et ont élaboré un modèle décomposant les représentations d'un même document en une partie informative et une partie résiduelle nuisible, représentée dans un espace de faible dimension. Les expérimentations menées avec les méthodes issues de l'analyse factorielle pour la réduction des représentation multi-vues montrent un gain substantiel en termes de bonne catégorisation de documents fortement bruités.

4.1 Introduction

Le chapitre 3 a présenté les avantages d'une représentation multiple lors d'une tâche de catégorisation d'un document bruité s'appuyant sur le modèle LDA dans des espaces de tailles (ou granularités) différentes. Ces différentes "vues" d'un même document permettent d'améliorer la robustesse des systèmes de catégorisation aux erreurs du système de reconnaissance de la parole.

Une des principales faiblesses du modèle LDA est le choix des valeurs des hyper-paramètres ainsi que du nombre de thèmes composant l'espace de thèmes, déterminant la granularité du modèle. Les performances des systèmes utilisant les espaces LDA peuvent être très instables si le processus de décision est fortement lié à ces hyper-paramètres.

De manière classique, cette représentation abstraite implique de choisir le nombre de thèmes composant l'espace de thèmes (K) aussi bien que les valeurs des hyper-paramètres (α et β) du modèle LDA. Les hyper-paramètres α et β contrôlent respectivement la distribution des thèmes au sein des documents et la distribution des termes dans chacun des thèmes de l'espace LDA. Le nombre de thèmes K contrôle, lui, la "granularité" du modèle, qui va d'une représentation thématique assez générale (peu de thèmes dans le modèle) à une représentation relativement précise (grand nombre de thèmes). Trouver les valeurs optimales de ces paramètres est crucial sachant que la perplexité du modèle, qui est une image de la qualité de l'espace thématique, est fortement dépendante de ces caractéristiques. De plus, le contexte multi-thèmes et la proximité de ces thèmes impliquent la nécessité de trouver une représentation plus complexe du document (Morchid et al., 2014b).

Dans ce chapitre, nous proposons une solution à ces deux problèmes en utilisant plusieurs espaces de thèmes obtenus en variant les hyper-paramètres du modèle LDA α, β ainsi que le nombre de classes n dans l'espace. Chacun de ces espaces produit une vue spécifique du document et notre objectif est d'extraire une information complémentaire et pertinente depuis ce grand nombre de vues différentes. Une difficulté potentielle inhérente à un aussi grand nombre de vues est liée à la diversité de celles-ci. En effet, ces vues introduisent une variabilité utile nécessaire pour la bonne représentation des différents contextes du document ainsi qu'une variabilité nuisible liée aux projections multiples dans des espaces de thèmes différents.

Le paradigme de l'Analyse Factorielle Jointe ou *Joint Factor Analysis* (JFA) considère que les multiples variabilités sont indépendantes. Ainsi la représentation dans un espace JFA (Kenny et al., 2008) permet de compenser la variabilité nuisible à l'intérieur de la session d'un même locuteur. Cette représentation est une extension du modèle GMM-*Universal Background Model* (UBM) (Reynolds et Rose, 1995). Dans (Dehak et al., 2011), les auteurs proposent d'extraire, depuis le super-vecteur GMM, une représentation compacte appelée i-vecteur (i pour la tâche d'$i$dentification du locuteur). L'objectif du processus de compression (extraction des i-vecteurs) est de représenter la variabilité du super-vecteur dans un espace de faible dimension appelé espace de variabilité totale.

Cette représentation en *i*-vecteurs, initialement introduite pour la vérification du locuteur (Kenny et al., 2008), est devenue très populaire dans le domaine du traitement automatique de la parole. Des publications récentes montrent que cette représentation est également fiable pour la reconnaissance de la langue (Martınez et al., 2011) ainsi que pour la segmentation en locuteurs (Franco-Pedroso et al., 2010). La réduction des représentations d'un document pour générer des *i*-vecteurs est une manière élégante de réduire la grande dimension des vecteurs d'entrée pour obtenir une représentation de faible dimension tout en conservant l'essentiel de l'information pertinente (voir partie 4.4).

Dans ce chapitre, nous proposons alors de réduire la variabilité inter-modèles en compactant les multiples vues d'un document en utilisant la méthode de l'analyse factorielle fournissant une représentation compacte, appelée *i*-vecteur. L'analyse factorielle est une méthode d'analyse de données très ancienne qui a été appliquée avec succès dans un premier temps à la tâche de vérification du locuteur, et par la suite généralisée à des tâches du domaine de la parole.

La partie 4.2 de ce chapitre présente les travaux précédents sur l'utilisation de l'analyse factorielle dans le domaine de la parole. La partie 4.3 présente la méthode d'extraction des *i*-vecteurs, ainsi que leur application dans le contexte de représentation multiple de documents bruités. La partie 4.4 introduit nos contributions portant sur l'application des *i*-vecteurs afin d'obtenir une représentation compacte et robuste de documents fortement bruités. Une conclusion sur l'apport de telles méthodes est ensuite proposée dans la partie 4.5.

4.2 Domaines d'application de l'analyse factorielle

La modélisation fondée sur l'analyse factorielle est appliquée, pour la première fois, dans le domaine de la vérification du locuteur (Kenny et al., 2005; Vogt et al., 2005; Matrouf et al., 2007). La partie 4.2.1 présente l'analyse factorielle dans le domaine de la Vérification du Locuteur (VL) ou *Speaker Recognition*. Dans un système de vérification du locuteur, cette modélisation permet d'améliorer la robustesse des systèmes face à la variabilité liée à la session [1] représentant une cause majeure de dégradation des performances du système. Elle est également couramment utilisée dans les domaines de la segmentation en locuteurs (voir partie 4.2.2), dans la reconnaissance de la parole (voir partie 4.2.3), ou bien encore le traitement de l'image (voir partie 4.2.4).

1. D'une manière générale, les termes *décalage de session* ou *variabilité session* sont utilisés pour désigner le changement des conditions acoustiques, qui peuvent varier grandement d'une session à l'autre. Le terme *variabilité session* englobe un grand nombre de phénomènes : le canal de transmission, le bruit environnant, la position du microphone, ...

4.2.1 L'analyse factorielle pour la vérification du locuteur

Le domaine de la vérification du locuteur (VL) a connu un essor important depuis l'utilisation du modèle fondé sur les mélanges de gaussiennes (Reynolds et Rose, 1995; Reynolds et al., 2000; Doddington et al., 2000). Ce modèle génératif est à l'état-de-l'art dans le domaine de la VL et est associé à des méthodes de classification telles que les machines à vecteurs de support (SVM). Cette méthode, dans le domaine de la catégorisation, est utilisée soit en combinaison des modèles génératifs (Dong et Zhaohui, 2001; Wan et Renals, 2003, 2005; Campbell et al., 2006; Dehak et Chollet, 2006; Dehak et al., 2007), soit directement sur les données acoustiques (Schmidt et Gish, 1996; Campbell, 2002).

Ainsi, les auteurs dans (Dong et Zhaohui, 2001) utilisent, comme entrées, les probabilités issues de modèles s'appuyant sur les GMM, à des classifieurs discriminants (SVM). Cette méthode est appliquée sur le corpus YOHO (Campbell et Higgins, 1994). Les premiers résultats encourageants obtenus ont conduit à l'application de cette combinaison (GMM+SVM) dans la tâche de vérification du locuteur. La méthode la plus performante et la plus utilisée, combinant le modèle génératif (GMM) ainsi que la méthode de classification (SVM), est l'association des deux paradigmes dans un système unique (Campbell et al., 2006; Dehak et Chollet, 2006; Dehak et al., 2007). Ce système utilise un noyau de distance probabiliste issu de la divergence de Kullback-Leibler (Kullback, 1987) entre les GMM. Avec cette méthode, l'entrée de l'espace des SVM coïncide avec la moyenne des GMM.

Le second ensemble de méthodes combinant les GMM et les classifieurs SVM utilise directement les représentations acoustiques du signal de parole. (Schmidt et Gish, 1996) utilisent lors de la phase d'apprentissage des SVM, les vecteurs acoustiques caractérisant les clients ainsi que les imposteurs. Durant la phase de test, le score du segment est obtenu en calculant la moyenne des scores des sorties du SVM pour chacune des trames. D'autres applications des SVM, dans le domaine de la vérification du locuteur existent. (Campbell, 2002) proposent la séquence discriminante linéaire généralisée ou *generalized linear discriminant sequence* (GLDS). Ce noyau, étant considéré comme le plus utilisé, offre l'avantage d'éliminer la variabilité liée au contexte, en moyennant les caractéristiques sur l'ensemble du vecteur de représentation.

Le principe d'une projection des données dans un espace de représentation commun uniforme fondé sur les mélanges de gaussiennes et de l'analyse factorielle s'est rapidement généralisé dans le domaine de la reconnaissance du locuteur, porté par les succés dans la tâche de vérification du locuteur (*Speaker Recognition Evaluation* (SRE)) des campagnes NIST (*National Institute of Standards and Technology*) [2] (Kenny et Dumouchel, 2004; Kenny, 2010).

Une des faiblesses que partage la méthode d'analyse factorielle (FA) avec d'autres modèles génératifs est le temps nécessaire à l'apprentissage du modèle. En effet, l'analyse factorielle, dans le cas de la vérification du locuteur, nécessite de présenter au système un grand nombre de contextes pour chacun des locuteurs lors de la phase

2. **http://www.nist.gov/itl/iad/mig/sre12.cfm**

d'apprentissage. Ainsi, pour pallier cette difficulté (Kenny et al., 2005) proposent d'utiliser une version simplifiée du modèle bâti sur l'analyse factorielle, en procédant à des approximations. Ceci est réalisé en observant un coût moindre en termes de perte de performance ($\leq 0,1$ %), tout en observant un gain substantiel en termes de temps de traitement (2 à 3 fois plus rapide). Cette implémentation permet l'utilisation de corpus de grandes tailles lors de la phase d'apprentissage. Le nombre de données ainsi que la diversité des données représentatives d'un même locuteur, sont primordiales pour obtenir de bonnes performances en termes de bonne vérification de locuteurs dans des conditions fortement bruitées. D'autres méthodes ont réduit le temps de calcul en réalisant, par exemple, un certain nombre d'hypothèses simplificatrices comme dans (Glembek et al., 2009). Dans (Kenny et al., 2005), les auteurs constatent un gain de temps de traitement important (jusqu'à près de 100 fois plus rapide) avec ce type d'approches.

Les méthodes d'analyse factorielle, utilisées dans ce chapitre, sont issues du cadre de la JFA et reposent sur le modèle théorique décrit dans le domaine de la vérification du locuteur (Kenny et al., 2007).

4.2.2 L'analyse factorielle pour la segmentation en locuteurs

L'objectif de la segmentation en locuteurs (Gish et al., 1991; Beigi et Maes, 1998; Chen et Gopalakrishnan, 1998; Siegler et al., 1997) est de diviser un flux audio en des segments homogènes. Ces segments doivent appartenir au même locuteur et être d'une durée la plus longue possible. La phase de segmentation en locuteurs est souvent en amont de tâches de plus haut niveau telles que le regroupement automatique de messages (Reynolds et al., 1998), le suivi de locuteurs (Rosenberg et al., 1998; Magrin-Chagnolleau et al., 1999; Mami et Charlet, 2002), l'indexation de locuteurs (Akita et Kawahara, 2003), ou encore la transcription automatique de journaux télévisés (Woodland et al., 1997; Gauvain et al., 1998). C'est un composant souvent essentiel dans la chaine de traitement de documents parlés.

La JFA pour la segmentation a été précédée par autre une approche qui présente quelques similitudes (Aronowitz, 2007)et qui consiste à projeter les segments de parole dans un espace de segments en utilisant un noyau ACP ou *kernel-PCA*. La projection dans l'espace de représentation commun est réalisée en estimant une distribution pour chacun des locuteurs, typiquement une distribution normale multivariée (PCA). Toutes les représentations probabilistes des locuteurs dans l'espace commun partagent la même matrice de covariance Σ. Les résultats encourageants obtenus par la modélisation de la variabilité inter-session intra-locuteur par un *kernel-PCA*, ont poussé les auteurs à évaluer le paradigme de la JFA dans le domaine de la segmentation et du regroupement en locuteurs. Ainsi, dans (Aronowitz, 2010), les auteurs, sont parmi les premiers à employer un système utilisant la JFA pour la segmentation en locuteurs. La méthode proposée ne nécessite pas de phase d'apprentissage, mais elle estime et compense la variabilité inter-session intra-locuteur "à la volée". Les performances enregistrées permettent de réduire de plus de 42 % le SER (*Speaker Error Rate*) sur le corpus d'évaluation de la campagne NIST-2005 SRE (NIST, 2005). Les travaux de (Kenny,

2008, 2010; Zhao et Dong, 2012) proposent de combiner l'analyse factorielle et l'approche naïve bayésienne. Cette approche permet d'améliorer encore les performances observées dans la tâche de segmentation et de regroupement en locuteurs.

4.2.3 L'analyse factorielle pour la reconnaissance de la parole

Le domaine de la reconnaissance de la parole fait également appel à la JFA pour transcrire au mieux des documents parlés. Une des méthodes issue de l'analyse factorielle est la projection des données dans un espace de représentation commun appelé *Subspace GMM* (Modèle de sous-espace de mélanges de gaussiennes ou *Subspace Gaussian Mixture Model* (SGMM) (Povey et al., 2010)). Les auteurs proposent de réduire l'espace de représentation des vecteurs associés aux états afin de contrôler des paramètres du GMM partagé entre tous les états (moyennes et mélanges de poids), par l'intermédiaire des paramètres communs partagés par tous les états. Chacun des segments de parole est lui-même un mélange de sous-états.

Une approche très semblable a été proposée dans (Bouallegue et al., 2011, 2012) et appliquée à la modélisation acoustique compacte. La méthode proposée consiste à représenter l'espace acoustique de la parole incluant tous les phonèmes par un seul modèle générique (GMM-UBM), puis à dériver les modèles des différents états de MMC depuis ce modèle générique, en mutualisant une large partie des modèles. Les dérivations des moyennes des gaussiennes sont obtenues en utilisant un modèle fondé sur l'analyse factorielle, alors que les moyennes initiales sont obtenues en utilisant le maximum de vraisemblance. Les variances sont restées inchangées par rapport au l'UBM. Un intérêt indéniable de cette représentation vectorielle est la possibilité de traiter les états par des techniques classiques d'*analyse de données* ou de classification automatique.

De nombreuses techniques ont été proposées pour augmenter la robustesse des systèmes de TRAP, en particulier leurs résistances aux bruits. L'objectif de ces techniques est de compenser les différences entre les conditions d'apprentissage et les conditions d'utilisation du système. L'analyse factorielle a été utilisée aussi dans ce domaine. (Rouvier et al., 2011; Bouallegue et al., 2012) proposent une nouvelle méthode de compensation des multiples variabilités nuisibles pour les systèmes de TRAP. L'estimation et l'isolation du bruit se fait en utilisant l'analyse factorielle dans l'espace des *super-vecteurs* formé par la concaténation des moyennes des gaussiennes composant le GMM-UBM. Les auteurs supposent que les variabilités nuisibles sont additives dans ce domaine et qu'elles sont situées dans un sous-espace de très faible dimension (comparativement à la dimension du super-vecteur). Les résultats obtenus sur la transcription de la parole bruité ont montré l'efficacité de cette méthode de filtrage basée sur l'analyse factorielle.

4.2.4 L'analyse factorielle dans le domaine du traitement d'image

L'apport de l'analyse factorielle dans le domaine de la parole a éveillé la curiosité des chercheurs d'autres domaines confrontés à la même problématique : représenter et

compenser la variabilité induite dans les différentes représentations d'un même individu d'une classe donnée.

Les premiers à avoir utilisé la JFA dans le domaine du traitement de l'image (Vesnicer et al., 2012) ont montré qu'une version simplifiée de la *Probabilistic Linear Discriminant Analysis* (PLDA) (Li et al., 2012) permet d'obtenir les meilleurs résultats que les méthodes fondées sur la PLDA ou l'ACP dans la tâche de reconnaissance faciale sur le corpus FRGCv2 (Li et al., 2012). L'analyse factorielle a par ailleurs été appliquée avec succès à l'identification du genre vidéo dans (Rouvier et al., 2009; Matrouf et al., 2011).

4.3 Représentation compacte au moyen d'un *i*-vecteur

La méthode des *i*-vecteurs utilisée pour compacter différentes vues d'un même document, est issue de l'analyse factorielle. La partie suivante 4.3.1 décrit le contexte d'utilisation des *i*-vecteurs dans le domaine de la reconnaissance du locuteur. La partie 4.3.2 décrit la procédure d'estimation des *i*-vecteurs dans le cas de multiples représentations d'un même document dans des espaces de thèmes issus de LDA.

4.3.1 Définition de l'espace de variabilité totale pour l'élaboration des *i*-vecteurs

L'extraction des *i*-vecteurs peut être vue comme un processus de compression permettant de réduire la dimensionnalité du super-vecteur de parole sachant un modèle linéaire Gaussien. (h, s) indique la session h du locuteur s. Le super-vecteur de parole $\mathbf{m}_{(h,s)}$ des moyennes des GMM concaténées d'un enregistrement de parole donné, est projeté dans un espace de faible dimension appelé espace de variabilité totale :

$$\mathbf{m}_{(h,s)} = \mathbf{m} + \mathbf{U}\mathbf{x}_{(h,s)} \tag{4.1}$$

où $\mathbf{m}$ est le super-vecteur moyen de l'UBM [3]. $\mathbf{U}$ est une matrice de rang faible et de dimension $MD \times R$, où M est le nombre de gaussiennes contenues dans l'UBM et D est le nombre de caractéristiques cepstrales représentant la base de l'espace réduit de variabilité totale. La matrice $\mathbf{U}$ est appelée *matrice de variabilité totale*. Les composantes de $\mathbf{x}_{(h,s)}$ sont les facteurs totaux représentant les coordonnées de l'enregistrement de parole dans l'espace réduit de variabilité totale appelés *i*-vecteurs.

4.3.2 Du *i*-vecteur pour la vérification du locuteur au *c*-vecteur pour la catégorisation de documents

L'approche proposée utilise les *i*-vecteurs pour modéliser les représentations des documents textuels à travers chacun des espaces de thèmes dans un espace de vocabulaire homogène. Ce segment court est considéré comme la représentation unitaire

3. L'UBM est un GMM représentant toutes les observations possibles.

s'appuyant sur le contenu sémantique du document. En effet, un vecteur de représentation d'un document dans un espace de vocabulaire correspond à un segment d'un document d. Dans la suite, (d,r) indique la représentation du document d dans l'espace de thèmes LDA r. Dans le modèle proposé dans la figure 4.1, le super-vecteur $\mathbf{m}_{(d,r)}$ d'un document d sachant l'espace de thèmes r est modélisé ainsi :

$$\mathbf{m}_{(d,r)} = \mathbf{m} + \mathbf{U}\mathbf{x}_{(d,r)} \tag{4.2}$$

où $\mathbf{x}_{(d,r)}$ sont les coordonnées de la représentation fondée sur les espaces de thèmes du document d dans l'espace réduit de variabilité totale appelé c-vecteur (c pour *c*atégorisation).

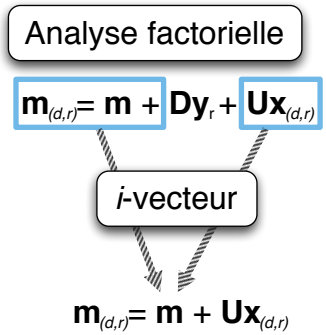


FIGURE 4.1 – *Composantes utiles pour l'extraction des i-vecteurs.*

$\mathbf{N}_{(d,r)}$ et $\mathbf{X}_{(d,r)}$ sont les deux vecteurs contenant les statistiques d'ordre 0 et d'ordre 1 du document d. Les statistiques sont estimées en utilisant le GMM-UBM :

$$\mathbf{N}_r[g] = \sum_{t \in r} \gamma_g(t);\ \{\mathbf{X}_{(d,r)}\}_{[g]} = \sum_{t \in (d,r)} \gamma_g(t) \cdot t \tag{4.3}$$

où $\gamma_g(t)$ est la probabilité *a posteriori* de la gaussienne g pour l'observation t. Dans l'équation, $\sum_{t \in (d,r)}$ représente la somme sur toutes les trames appartenant au dialogue d.

$\overline{\mathbf{X}}_{(d,r)}$ sont les statistiques dépendant de l'état défini comme suit :

$$\{\overline{\mathbf{X}}_{(d,r)}\}_{[g]} = \{\mathbf{X}_{(d,r)}\}_{[g]} - \mathbf{m}_{[g]} \cdot \sum_{(d,r)} \mathbf{N}_{(d,r)}[g] \tag{4.4}$$

Soit $\mathbf{L}_{(d,r)}$ une matrice $R \times R$ et $\mathbf{B}_{(d,r)}$ un vecteur de dimension R, les deux étant définis comme suit :

$$\begin{aligned} \mathbf{L}_{(d,r)} &= \mathbf{I} + \sum_{g \in \text{UBM}} \mathbf{N}_{(d,r)}[g] \cdot \{\mathbf{U}\}^t_{[g]} \cdot \Sigma^{-1}_{[g]} \cdot \{\mathbf{U}\}_{[g]} \\ \mathbf{B}_{(d,r)} &= \sum_{g \in \text{UBM}} \{\mathbf{U}\}^t_{[g]} \cdot \Sigma^{-1}_{g} \cdot \{\overline{\mathbf{X}}_{(d,r)}\}_{[g]}, \end{aligned} \tag{4.5}$$

Algorithm 3: Algorithme d'estimation de $\mathbf{U}$ ainsi que de la variable latente $\mathbf{x}_{(d,r)}$.

Pour chaque document d projeté dans l'espace de thèmes r : $\mathbf{x}_{(d,r)} \leftarrow 0$,
$\mathbf{U} \leftarrow$ *aleatoire* ;
Estimation des statistiques : $\mathbf{N}_{(d,r)}$, $\mathbf{X}_{(d,r)}$ (eq. 4.3);
for $i = 1$ *to nb_iterations* **do**
  for *tous les documents d et espaces de thèmes r* **do**
    Statistiques centrées : $\overline{\mathbf{X}}_{(d,r)}$ (eq. 4.4);
    Estimation de $\mathbf{L}_{(d,r)}$ et $\mathbf{B}_{(d,r)}$ (eq. 4.5);
    Estimation de $\mathbf{x}_{(d,r)}$ (eq. 4.6);
  end
  Estimation de la matrice $\mathbf{U}$ (eq. 4.7 and 4.8) ;
end

En utilisant $\mathbf{L}_{(d,r)}$ ainsi que $\mathbf{B}_{(d,r)}$, $\mathbf{x}_{(d,r)}$ peut alors être obtenu en utilisant l'équation suivante :

$$\mathbf{x}_{(d,r)} = \mathbf{L}_{(d,r)}^{-1} \cdot \mathbf{B}_{(d,r)} \tag{4.6}$$

La matrice $\mathbf{U}$ peut être estimée ligne par ligne, avec $\{\mathbf{U}\}_{[g]}^{i}$ étant la $i^{\text{ème}}$ ligne de $\{\mathbf{U}\}_{[g]}$ alors :

$$\mathbf{U}_{[g]}^{i} = \mathcal{L}_g^{-1} \cdot \mathcal{R}_g^{i}, \tag{4.7}$$

où $\mathcal{R}_g^{i}$ et $\mathcal{L}_g$ sont donnés par :

$$\begin{aligned} \mathcal{L}_g &= \sum_{(d,r)} \mathbf{L}_{(d,r)}^{-1} + \mathbf{x}_{(d,r)} \mathbf{x}_{(d,r)}^{t} \cdot \mathbf{N}_{(d,r)}[g] \\ \mathcal{R}_g^{i} &= \sum_{(d,r)} \{\overline{\mathbf{X}}_{(d,r)}\}_{[g]}^{[i]} \cdot \mathbf{x}_{(d,r)} \end{aligned} \tag{4.8}$$

L'algorithme 4 présente la méthode issue de l'algorithme original 4 et adoptée pour l'estimation de la matrice de variabilité des représentations multi-vues des documents fortement bruités issus de transcriptions automatiques. La fonction standard de vraisemblance peut être utilisée pour évaluer la convergence de l'algorithme.

La représentation en c-vecteur souffre néanmoins de trois problèmes :

- Le c-vecteur $\mathbf{x}_{(d,r)}$ de l'équation 4.2 doit théoriquement être distribué normalement sur la distribution normale $\mathcal{N}(0, I)$,
- l'effet "radial" doit être supprimé,
- le rang de l'espace des facteurs totaux doit être utilisé pour procéder à des transformations discriminantes.

Une solution bâtie sur la standardisation des données est proposée dans la partie 4.4.4. Avant cela, la partie qui suit présente la représentation multiple fondée sur des espaces de thèmes LDA différents.

4.4 Contributions : Représentation compacte de documents bruités s'appuyant sur l'espace des *i*-vecteurs

La partie précédente décrit plusieurs domaines faisant appel à l'analyse factorielle pour compenser un bruit résiduel, pour réduire l'espace de représentation, ... Le chapitre 4 montre l'apport d'une représentation multiple d'un document dans divers espaces thématiques. Cette représentation multiple de documents bruités tels des transcriptions automatiques, introduit deux variabilités, l'une utile, correspondant aux différentes vues d'un même document, ainsi qu'une autre variabilité résiduelle correspondant aux erreurs de transcription et aux projections multiples. Cette partie décrit nos contributions pour pallier, ou du moins compenser, cette variabilité dite "résiduelle" dans le domaine du traitement de documents bruités et d'articles de journaux.

4.4.1 Représentation des documents bruités dans un espace de vocabulaire homogène

Cette partie décrit le processus de projection multiple d'un document dans des espaces de thèmes LDA. Un ensemble de caractéristiques permettant de capturer les relations statistiques entre les mots composant le document est obtenu par une allocation latente de Dirichlet (LDA).

Pour un ensemble de documents composant le corpus d'apprentissage D, un ensemble d'espaces de thèmes est appris et un document d est représenté par une distribution de probabilités dans chacun des espaces de thèmes cachés. L'estimation de ces probabilités est affectée par une variabilité inhérente à l'estimation des hyper-paramètres du modèle. Si plusieurs espaces sont considérés et les caractéristiques sont déterminées pour chacun des espaces LDA, il est alors possible de modéliser la variabilité utile et la variabilité liée aux expressions linguistiques utilisées pour exprimer un même thème, par des locuteurs différents, dans une situation réelle.

Plusieurs techniques, comme les méthodes variationnelles (Blei et al., 2003), de propagation EM (Minka et Lafferty, 2002) ou le *Gibbs Sampling* (Griffiths et Steyvers, 2004) détaillées dans la partie 1.3.3.2, ont été proposées pour estimer les paramètres du modèle LDA (α, β, et le nombre de thèmes composant le modèle). Cette dernière méthode d'estimation des hyper-paramètres du modèle LDA est un cas particulier des MCMC (Geman et Geman, 1984) (*Markov-chain Monte Carlo*). Elle fournit un algorithme simple d'approximation des paramètres dans un espace de grande dimension comme LDA (Heinrich, 2005). Ceci permet de contourner la difficulté liée à une estimation directe et précise des paramètres maximisant la vraisemblance de l'ensemble des données définie ainsi :

$$P(W|\alpha,\beta) = \prod_{w \in W} P(\overrightarrow{w}|\alpha,\beta) \tag{4.9}$$

pour toutes les données de la collection W connaissant les paramètres $\overrightarrow{\alpha}$ et $\overrightarrow{\beta}$. Ainsi, à la fin de l'algorithme LDA, un modèle contenant K classes qui contiennent elles-mêmes

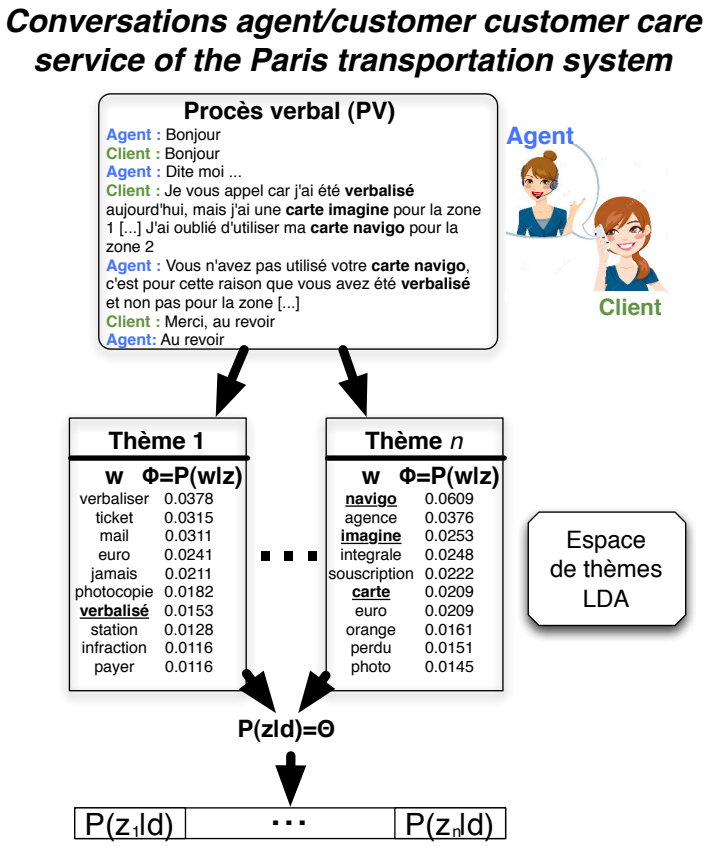


FIGURE 4.2 – *Projection d'un document dans un espace de thèmes LDA.*

une distribution $\theta_{w,z}$ est obtenu. Cette distribution de probabilités qu'un terme w du vocabulaire du corpus d'apprentissage soit généré par la classe z est calculée comme suit :

$$\theta_{w,z} = P(w|z) \; . \tag{4.10}$$

4.4.2 Variation des paramètres du modèle LDA pour une représentation multi-vues d'un document

Un ensemble de p espaces de thèmes (ici, $p = 500$) sont appris en utilisant LDA en faisant varier les hyper-paramètres :

- Le nombre de thèmes K dans l'espace de thèmes (partie 4.4.2.1),
- l'hyper-paramètre α (partie 4.4.2.2),
- et l'hyper-paramètre β (partie 4.4.2.3).

4.4.2.1 Variation du nombre de thèmes K

Le nombre de classes varie pour obtenir un ensemble de p espaces de taille K ($5 \leq K \leq 505$). En tout, 500 espaces thématiques sont estimés. Ce nombre est assez élevé pour générer, pour chacun des documents, un nombre d'observations suffisant pour estimer les paramètres du modèle de variabilité totale.

4.4.2.2 Variation de α

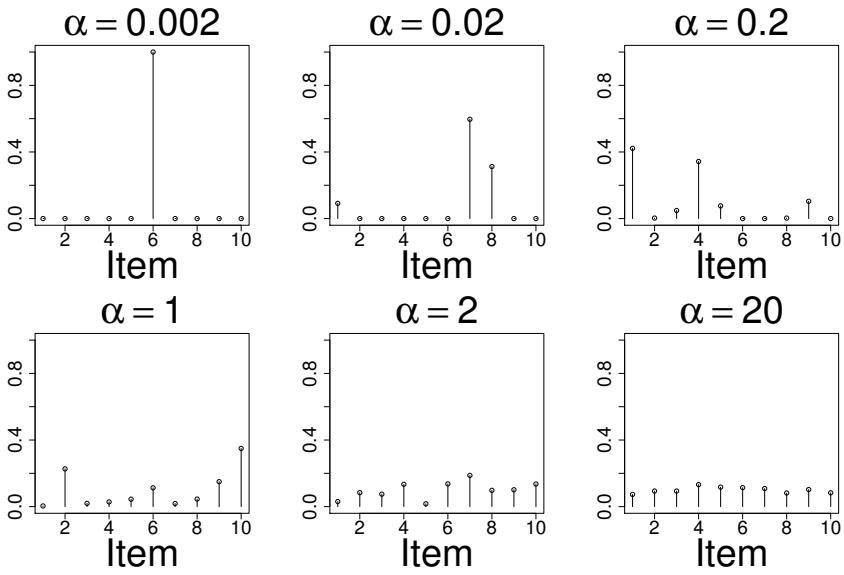


FIGURE 4.3 – *Distribution de la loi de Dirichlet pour différentes valeurs de α.*

Durant le processus d'apprentissage du modèle LDA, un thème z est tiré depuis une loi multinomiale de paramètre θ qui est elle-même tirée depuis une distribution de Dirichlet de paramètre $\overrightarrow{\alpha}$. Ainsi, un ensemble de p espaces de thèmes de taille K est appris en faisant varier le paramètre contrôlant la distribution des thèmes $\overrightarrow{\alpha}$.

L'heuristique standard est $\alpha_0 = \frac{50}{K}$ (Griffiths et Steyvers, 2004), qui correspond pour le $i^{\text{ème}}$ espace thématique ($1 \leq i \leq p$) à $\overrightarrow{\alpha_i} [\underbrace{\alpha_i, \ldots, \alpha_i}_{\text{K fois}}]^t$ avec :

$$\begin{aligned} \alpha_i &= \frac{i}{p} \times \alpha_0 \\ &= \frac{i}{p} \times \frac{50}{K} . \end{aligned} \tag{4.11}$$

Plus la valeur de α_i est grande ($\alpha_i \geq 1$), plus la distribution $P(z|d)$ sera uniforme (voir figure 4.3 avec $\alpha = 20$) et un document d sera associé avec des probabilités $P(z|d)$

proches de la plupart des thèmes z du modèle. Cependant, le but recherché est différent : des documents différents doivent être associés à différentes distributions $P(z|d)$. Dans le même temps, si la valeur de α est faible, plus les valeurs tirées selon une distribution de Dirichlet seront concentrées autour d'un nombre réduit de thèmes z (voir figure 4.3 avec $\alpha = 0.002$). Le nombre de thèmes K est fixé à 50 et le nombre d'espaces de thèmes est lui fixé à 500 ($p = 500$) durant nos expériences. Ceci permet de faire varier α_K d'une valeur faible (distribution $P(z|d)$ creuse avec $\alpha_1 = 0.002$) à 1 (distribution uniforme Dirichlet $\alpha_p = 1$).

4.4.2.3 Variation de β

De la même manière, l'hyper-paramètre β contrôle la distribution des termes dans chacun des thèmes de l'espace LDA. Plus la valeur de β est élevée, plus la distribution $P(w|z)$ est uniforme et sera semblable pour l'ensemble des thèmes composant l'espace LDA. Ceci implique que les classes du modèle sont elles-mêmes thématiquement proches. Durant la phase d'inférence permettant de représenter un document dans l'espace LDA, la distribution des thèmes pour un document donné se doit d'être différente.

La figure 4.3 représente des distributions de Dirichlet avec des valeurs de α différentes, pouvant également s'appliquer à l'hyper-paramètre β. L'heuristique standard pour le choix de la valeur de β est $\beta_0 = 0,1$ (Griffiths et Steyvers, 2004), ce qui correspond pour le $i^{\text{ème}}$ espace LDA ($1 \leq i \leq p$) à $\overrightarrow{\beta_i}[\underbrace{\beta_i, \ldots, \beta_i}_{|V| \text{times}}]^t$ avec :

$$\begin{aligned} \beta_i &= \frac{i}{p} \times \beta_0 \\ &= \frac{i}{p} \times 0,1 \,. \end{aligned} \tag{4.12}$$

Tout comme pour le paramètre α, le nombre de classes K est fixé à 50 et le nombre d'espaces de thèmes p est lui fixé à 500 ($p = 500$) durant nos expérimentations. Ainsi, β_i varie entre une valeur faible (représentation creuse des termes au sein des thèmes $\beta_1 = 0.0002$) à 1 (distribution uniforme de Dirichlet $\beta_p = 0.1$).

Le prochain processus nous permet d'obtenir une représentation homogène du document d pour le $i^{\text{ème}}$ espace de thèmes. La partie 4.4.3 présente la projection de chacune des représentations thématiques d'un document dans un espace homogène composé d'un vocabulaire discriminant V (Morchid et al., 2014b).

4.4.3 Représentation multiple dans un espace homogène de mots discriminants

Le vecteur de caractéristiques $V_d^{z^m}$ de d est projeté dans un espace commun composé des mots contenus dans V (Morchid et al., 2014b) de taille 166 dans nos expérimenta-

tions, pour obtenir un nouveau vecteur de caractéristiques $V_{d,i}^{w} = \{P(w|d)_i\}_{w \in V}$ de taille $|V|$ pour le $i^{\text{ème}}$ espace de thèmes de taille K avec la $j^{\text{ème}}$ $(0 \leq j \leq |V|)$ caractéristique :

$$\begin{aligned} V_{d,r}^{w_i} &= P(w_i|d) \\ &= \sum_{k=1}^{K} P(w_j|z_k^i) P(z_k^i|d) \\ &= \sum_{k=1}^{K} V_{z_k^i}^{w_j} \times V_d^{z_k^i} \\ &= \left\langle \overrightarrow{V_{z^i}^{w_j}}, \overrightarrow{V_d^{z^i}} \right\rangle \end{aligned}$$

où $\langle \cdot, \cdot \rangle$ représente le produit scalaire ; $V_{z_k^i}^{w_j} = P(w_i|z_k)$; $V_d^{z_k^i} = P(z_k|d)$ est estimé en utilisant le *Gibbs Sampling* pour le $i^{\text{ème}}$ espace de thèmes.

4.4.4 Standardisation des *c*-vecteurs

La fin de la partie 4.3 présente trois difficultés liées à une utilisation des méthodes fondées sur des GMM pour des données qui ne suivent pas forcément une loi gaussienne. Pour pallier ces difficultés, une transformation des représentations des documents bruités issus des ensembles d'apprentissage et de validation est réalisée pour la méthode *c*-vecteur issue de l'analyse factorielle. Selon (Bousquet et al., 2011), où les auteurs proposent pour la première fois l'algorithme de standardisation EFR, la première étape est d'évaluer la moyenne empirique $\overline{x}$ ainsi que la matrice de covariance $\mathbf{V}$ depuis les vecteurs d'apprentissage. La matrice de covariance $\mathbf{V}$ est décomposée en :

$$\mathbf{PDP}^T \,, \tag{4.13}$$

où $\mathbf{P}$ est la matrice des vecteurs propres de $\mathbf{V}$ et $\mathbf{D}$ est la version diagonale de $\mathbf{V}$. Un vecteur d'apprentissage $\mathbf{x}_{(d,r)}$ est transformé en $\mathbf{x}'_{(d,r)}$ comme suit :

$$\mathbf{x}'_{(d,r)} = \frac{\mathbf{D}^{-\frac{1}{2}} \mathbf{P}^t (\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})}{\sqrt{(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})^T \mathbf{V}^{-1} (\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})}} \tag{4.14}$$

Le numérateur est équivalent, par rotation, à $\mathbf{V}^{-\frac{1}{2}}(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})$, et la norme euclidienne de $\mathbf{x}'_{(d,r)}$ est égale à 1. La même transformation est appliquée aux vecteurs issus de l'ensemble de validation en utilisant l'ensemble des paramètres appris sur le corpus d'apprentissage $\overline{x}$ et la matrice de covariance $\mathbf{V}$ comme une estimation des paramètres de l'ensemble de validation. La figure 4.4 montre les différentes étapes :

- la figure 4.4-(a) présente l'ensemble original des vecteurs d'apprentissage,

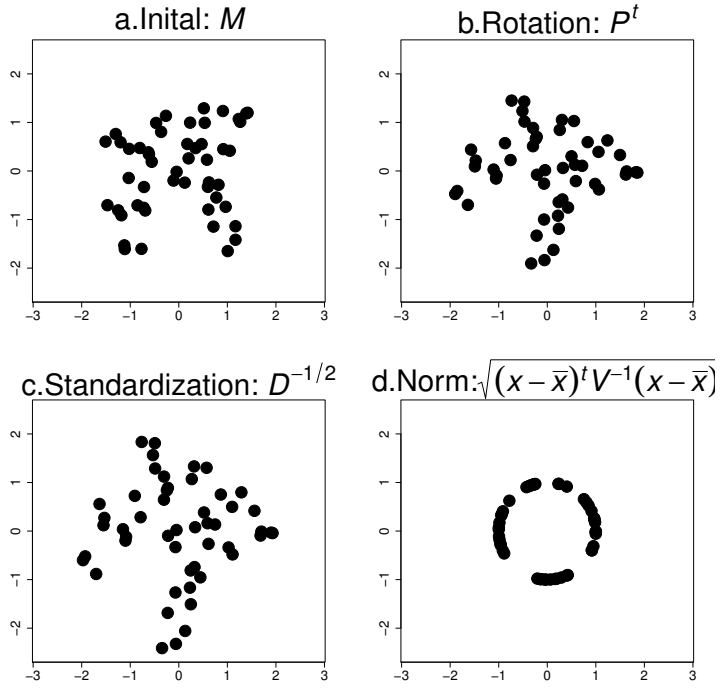


FIGURE 4.4 – *Effet de la standardisation avec l'algorithme EFR.*

- la figure 4.4-(b) montre l'effet de la rotation appliquée au données originales autour des axes de variabilité totale quand la transformation $\mathbf{P}^T$ est appliquée,
- la figure 4.4-(c) montre la standardisation du c-vecteur lorsque $\mathbf{D}^{-\frac{1}{2}}$ est appliqué,
- la figure 4.4-(d) présente le c-vecteur $\mathbf{x}'_{(d,r)}$ sur la surface de l'hyper-sphère de rayon 1 après la normalisation par la longueur du vecteur en divisant par $\sqrt{(\mathbf{x}_{(d,r)} - \bar{\mathbf{x}})^T \mathbf{V}^{-1} (\mathbf{x}_{(d,r)} - \bar{\mathbf{x}})}$.

4.4.5 Protocole expérimental

La représentation compacte proposée au moyen des c-vecteurs est évaluée dans le contexte de l'identification de thèmes dans des transcriptions automatiques de conversations téléphoniques composant le corpus DECODA et dans une tâche de catégorisation d'articles de presse issus du corpus Reuters-21578 ModApte (Asuncion et Newman, 2007). Cette représentation compacte est construite depuis un ensemble de vecteurs de représentation issus d'espaces LDA. La mesure de similarité permettant de catégoriser les documents est la distance de Mahalanobis décrite dans la partie 2.2.4.

La méthode LDA nous permet d'élaborer un ensemble de 500 espaces de thèmes en faisant varier le nombre de classes du modèle ou un des hyper-paramètres de ce même modèle (voir partie 4.4.2). Un espace de thèmes ayant moins de 5 classes n'est pas adapté à un corpus de grande taille comme celui utilisé durant cette évaluation. Comme

dans les parties précédentes, les espaces de thèmes LDA sont appris en utilisant l'outil Mallet[4].

Les parties suivantes décrivent le corpus Reuters ainsi que la métrique utilisée pour évaluer la pertinence des *c*-vecteurs pour la tâche de catégorisation des articles Reuters ainsi que de DECODA.

4.4.5.1 Corpus d'articles Reuters-21578

Pour évaluer la pertinence de la représentation compacte proposée dans ce chapitre, la tâche de catégorisation des 10 plus importantes classes en termes d'articles contenus Reuters-21578 ModApte (Asuncion et Newman, 2007) est utilisée. Le tableau 4.1 présente le nombre de documents pour la phase d'apprentissage, de *Dev* et de *Test* pour chacune des 10 classes (Gunal, 2012; Zhu et Lin, 2013).

TABLE *4.1* – *Top-10 des classes du corpus Reuters-21578.*

Classe ou étiquette	*Nombre d'articles*		
	apprentissage	**dev**	**test**
earn	2 590	287	1 087
acq	1 485	165	719
money-fx	484	54	179
grain	390	43	149
crude	350	39	189
trade	332	37	117
interest	312	35	131
ship	177	20	89
wheat	191	21	71
corn	163	18	56
Total	**6 474**	**719**	**2 787**

4.4.5.2 Mesure de similarité

La distance de Mahalanobis nous permet d'évaluer la similarité entre deux représentations vectorielles (ici, la représentation du document et le centroïde de la classe) et d'étiqueter le document. À la fin de ce processus, une métrique doit être choisie pour évaluer la performance du système de catégorisation proposé dans ce chapitre. Comme pour les chapitres précédents, la précision est choisie comme métrique d'évaluation de la tâche d'identification de thèmes dans le corpus de conversations DECODA.

La tâche de catégorisation d'articles Reuters est habituellement évaluée avec la Macro-F1. Pour cette raison, nous utilisons cette métrique pour évaluer la représentation compacte *c-vecteur*.

4. http ://mallet.cs.umass.edu/

Cette métrique est couramment utilisée dans le domaine de la catégorisation de documents. Elle est calculée pour chacune des classes dans le corpus et ensuite, la Macro-F1 moyenne entre toutes les classes est évaluée. Ainsi, un poids semblable est assigné à chacune des classes qui sont évaluées de la façon suivante :

$$\text{Macro-F1} = \sum_{k=1}^{K} \frac{F_k}{K} \ , \ F_k = \frac{2 \times p_k \times r_k}{p_k + r_k} \ , \tag{4.15}$$

où p_k et r_k sont respectivement la précision et le rappel pour la classe k parmi les K classes. Elles sont déterminées comme suit :

$$p_k = \frac{VP_k}{VP_k + FP_k} \quad \text{et} \quad r_k = \frac{VP_k}{VP_k + FN_k} \ . \tag{4.16}$$

FP_k représente le nombre de documents classifiés incorrectement dans la classe k (*i.e.* faux positifs) ; VP_k est le nombre de documents correctement classifiés dans la classe k (*i.e.* vrais positifs) ; FN_k représente le nombre de documents appartenant à la classe k mais n'étant pas classifiés dans cette classe (*i.e.* faux négatifs). Plus de détails concernant les métriques d'évaluation sont disponibles dans (Van Asch, 2013).

4.4.6 Résultats

TABLE 4.2 – *Configurations pour la représentation multi-granulaires.*

	Configuration des hyper-paramètres		
	K	α	β
K	$\mathbf{5} \leq K \leq \mathbf{505}$	$\alpha = \frac{50}{50} = 1$	$\beta = 0.1$
α	$K = 50$	$\mathbf{0.002} \leq \alpha \leq \mathbf{1}$	$\beta = 0.1$
β	$K = 50$	$\alpha = \frac{50}{50} = 1$	$\mathbf{0.0002} \leq \beta \leq \mathbf{0.1}$

Les expérimentations sont conduites en utilisant un ensemble d'espaces de thèmes estimés avec l'algorithme LDA. Les premières expérimentations présentées dans la partie 4.4.6.1 sont conduites sur la tâche d'identification de thèmes dans le corpus DECODA. La représentation compacte c-vecteur est ensuite évaluée dans le cadre de la catégorisation d'articles Reuters dans la partie 4.4.5.1.

Pour les deux tâches, des espaces de thèmes sont appris en faisant varier les valeurs des hyper-paramètres du modèle LDA ainsi que le nombre de classes comme décrit dans le tableau 4.2.

4.4.6.1 Représentation compacte de transcriptions automatiques bruitées

Les figures 4.5 et 4.6 présentent respectivement les performances obtenues lors de la tâche d'identification de thèmes du corpus DECODA pour les corpus de *Dev* et de *Test*

en utilisant différents configurations ainsi que l'algorithme EFR de normalisation pour l'ensemble de données TRAP (*i.e.* transcriptions automatiques) et TMAN (*i.e.* transcriptions manuelles) du corpus DECODA (*baseline*).

La première remarque est que les meilleures performances pour le système *baseline*, en termes de précision, sont obtenues lorsque l'on varie le paramètre α (contrôlant la distribution des thèmes pour un document) pour les sorties TRAP du corpus DECODA. Ces résultats atteignent 86,9 % et 80,1 % respectivement pour les ensembles de documents issus du *Dev* et du *Test*.

Ensuite, les résultats obtenus lorsque l'on varie l'hyper-paramètre β obtiennent la seconde meilleure précision en termes d'identification de thèmes avec 82,9 % et 74,0 % pour les ensembles de documents issus du *Dev* et du *Test*.

Nous pouvons noter que ces résultats sont proches de ceux obtenus avec une représentation multiple en faisant varier le nombre de classes n contenues dans l'espace de thèmes (81,7 % et 76,4 % pour les ensembles de *Dev* et de *Test*).

Sachant que la différence entre les résultats obtenus pour β et n pour le corpus de *Dev* dans la configuration TMAN sont importants ($90,3 - 88,6 = 1,7$ points), les résultats observés dans des conditions similaires sont proches (83,8 %). Ceci est également observé dans la configuration TRAP (74,0 % et 76,4 %).

Cependant, nous notons que la performance de catégorisation est très instable et peut complètement changer d'une configuration à l'autre. La différence entre la plus petite et la plus grande précision est également importante, avec une différence de 32,6 points pour le TRAP et 27,5 points pour le TMAN pour le *Dev* (la même tendance est observée pour le *Test*) lorsque le nombre de classes n varie. La recherche du meilleur nombre de classes paraît donc crucial pour la tâche de catégorisation, particulièrement dans le contexte de transcriptions automatiques bruitées.

On peut noter que les précisions les plus élevées sont obtenues lorsque l'hyper-paramètre α varie, alors que celles obtenues lorsque β et K varient sont presque similaires, ce qui n'est pas très intuitif. Normalement, le nombre de thèmes devrait avoir un impact plus important dans la représentation statistique du modèle LDA que l'hyper-paramètre α. Cependant, cette remarque est pertinente quand l'objectif est de projeter le document dans un espace unique. Ainsi, le processus d'inférence est sensible à la distribution des thèmes contrôlée par l'hyper-paramètre α pour un document n'apparaissant pas dans le corpus d'apprentissage dans l'espace de thèmes. Le sujet ici est la constitution de différentes représentations d'un même document pour contourner le problème d'un choix aveugle ou empirique des hyper-paramètres du modèle LDA et de considérer les différentes vues d'un même document.

Les tableaux 4.3 et 4.4 présentent les précisions obtenues avec la représentation compacte *c*-vecteur couplée avec l'algorithme de standardisation EFR, en prenant en considération différentes tailles pour les *c*-vecteurs et pour le nombre de gaussiennes contenues dans le GMM-UBM pour le corpus TMAN et TRAP. Les résultats présentés dans le tableau 4.5 montrent les précisions obtenues en termes d'identification de thèmes pour le corpus DECODA (TMAN/TRAP) avec la variation des hyper-paramètres LDA

TABLE 4.3 – *Précision (%) avec différentes tailles de c-vecteurs et de nombre de gaussiennes contenues dans le GMM-UBM pour l'ensemble d'apprentissage issu du TMAN → Test issu du TMAN.*

(c) Variation du nombre de thèmes K

taille du	*DEV*					*TEST*				
	Nombre de gaussiennes dans le GMM-UBM									
***c*-vecteur**	*32*	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	88.6	89.7	89.7	90.3	90.3	88.2	88.0	89.6	88.4	88.0
80	88.0	87.4	89.7	89.1	90.3	89.2	87.5	89.0	89.3	87.8
100	89.1	84.6	89.1	**92.0**	89.7	89.2	88.0	91.7	**89.3**	87.5
120	88.0	86.3	86.9	89.1	90.9	88.7	87.8	88.7	86.2	89.3
140	87.4	85.1	86.3	89.1	88.6	88.3	85.6	86.5	86.2	87.8
160	87.1	85.7	86.3	85.1	90.9	87.7	88.4	89.3	87.4	87.1
180	88.0	85.1	82.3	86.9	86.3	87.0	85.3	88.4	83.2	86.2
200	89.1	82.9	86.3	85.1	88.0	88.0	88.0	87.5	82.9	87.5

(a) Variation du paramètre α

taille du	*DEV*					*TEST*				
	Nombre de gaussiennes dans le GMM-UBM									
***c*-vecteur**	*32*	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	72.3	77.1	79.1	75.4	73.1	85.6	86.2	82.6	82.6	70.3
80	79.7	82.6	78.9	81.7	75.4	86.2	82.9	82.6	82.6	72.2
100	76.6	**88.3**	84.0	82.6	73.1	83.5	**85.6**	86.2	80.4	73.4
120	83.7	87.4	80.9	77.4	72.6	80.7	80.7	82.3	82.6	71.9
140	82.6	77.1	82.6	78.9	81.1	80.4	82.3	80.4	78.6	74.6
160	74.9	72.3	74.6	76.6	75.7	77.7	78.6	75.8	70.0	74.0
180	79.1	75.1	76.0	74.3	77.4	74.9	82.0	68.5	63.3	76.5
200	76.0	71.7	74.0	72.9	73.4	78.6	78.9	64.5	64.5	64.5

(b) Variation du paramètre β

taille du	*DEV*					*TEST*				
	Nombre de gaussiennes dans le GMM-UBM									
***c*-vecteur**	*32*	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	86.9	83.4	85.1	83.4	80.6	87.2	84.4	83.5	75.8	79.2
80	86.3	**87.4**	84.6	79.4	77.7	84.4	**84.7**	85.6	77.1	75.8
100	86.9	86.3	81.1	83.4	74.9	85.9	82.6	85.0	74.3	76.7
120	81.7	84.6	83.4	78.9	77.7	83.5	84.4	80.4	69.4	72.5
140	72.0	81.7	84.0	79.4	72.0	79.8	78.3	80.4	76.4	70.6
160	79.4	81.7	82.3	75.4	78.9	80.4	81.3	78.6	81.3	71.2
180	74.9	84.6	77.7	77.1	68.6	81.9	79.5	77.0	81.9	65.1
200	82.9	84.0	79.4	75.4	72.0	81.9	79.2	77.7	62.7	70.0

TABLE 4.4 – *Précision (%) avec différentes tailles de c-vecteurs et de nombre de gaussiennes contenues dans le GMM-UBM pour l'ensemble d'apprentissage issu du TRAP → Test issu du TRAP.*

(c) Variation du nombre de thèmes K

taille du *c*-vecteur	*DEV*					*TEST*				
	Nombre de gaussiennes dans le GMM-UBM									
	32	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	82.9	82.9	88.0	84.0	86.3	85.0	83.2	83.2	83.8	84.7
80	85.1	81.7	84.6	82.9	86.3	82.9	79.5	83.2	84.7	82.0
100	82.4	82.9	85.1	**89.7**	88.0	84.1	81.0	84.7	**86.5**	83.5
120	83.4	84.0	81.7	87.4	85.1	82.9	83.2	85.0	84.4	81.6
140	81.1	84.6	87.4	84.6	83.4	85.3	82.9	84.1	82.9	82.9
160	78.9	82.3	82.9	83.4	82.9	86.5	84.7	80.7	82.9	81.7
180	81.1	81.7	80.6	83.4	82.3	85.0	80.7	79.8	78.6	79.2
200	82.3	82.9	84.6	81.1	81.7	83.5	81.3	79.8	79.8	76.1

(a) Variation du paramètre α

taille du *c*-vecteur	*DEV*					*TEST*				
	Nombre de gaussiennes dans le GMM-UBM									
	32	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	81.1	82.9	78.9	78.9	77.1	82.9	78.6	72.8	80.1	71.0
80	82.9	84.6	75.4	80.6	76.0	81.3	79.2	73.4	78.3	69.1
100	85.7	**86.9**	79.4	80.0	72.6	78.9	**80.4**	80.1	75.5	63.9
120	80.0	82.9	75.5	74.3	73.7	83.2	80.4	71.0	70.6	60.2
140	79.4	80.0	79.4	76.0	68.6	74.3	73.1	75.0	69.4	59.6
160	74.3	80.6	73.1	78.3	68.0	77.7	75.2	70.3	68.2	61.2
180	76.6	83.4	72.0	73.7	66.3	72.2	71.9	61.2	66.1	57.5
200	72.0	74.3	68.0	66.9	67.4	70.3	70.3	68.8	66.4	55.4

(b) Variation du paramètre β

taille du *c*-vecteur	*DEV*					*TEST*				
	Nombre de gaussiennes dans le GMM-UBM									
	32	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	84.0	84.0	74.9	76.0	73.7	73.0	77.3	70.8	75.1	67.5
80	76.6	83.4	79.4	74.3	76.0	73.6	75.1	72.4	72.7	69.9
100	71.4	**85.7**	74.3	76.0	76.6	77.8	**75.4**	68.7	68.5	69.9
120	74.3	77.7	71.4	76.6	65.7	80.4	70.5	68.7	75.7	66.6
140	66.6	78.3	63.4	64.6	65.1	73.7	69.0	67.8	67.2	62.6
160	69.4	76.0	68.6	61.1	65.1	70.3	74.5	70.2	63.2	69.3
180	70.9	77.1	68.6	65.7	62.9	70.9	72.4	68.3	61.4	66.0
200	68.6	73.7	60.6	66.3	56.0	69.7	63.9	65.0	63.8	68.0

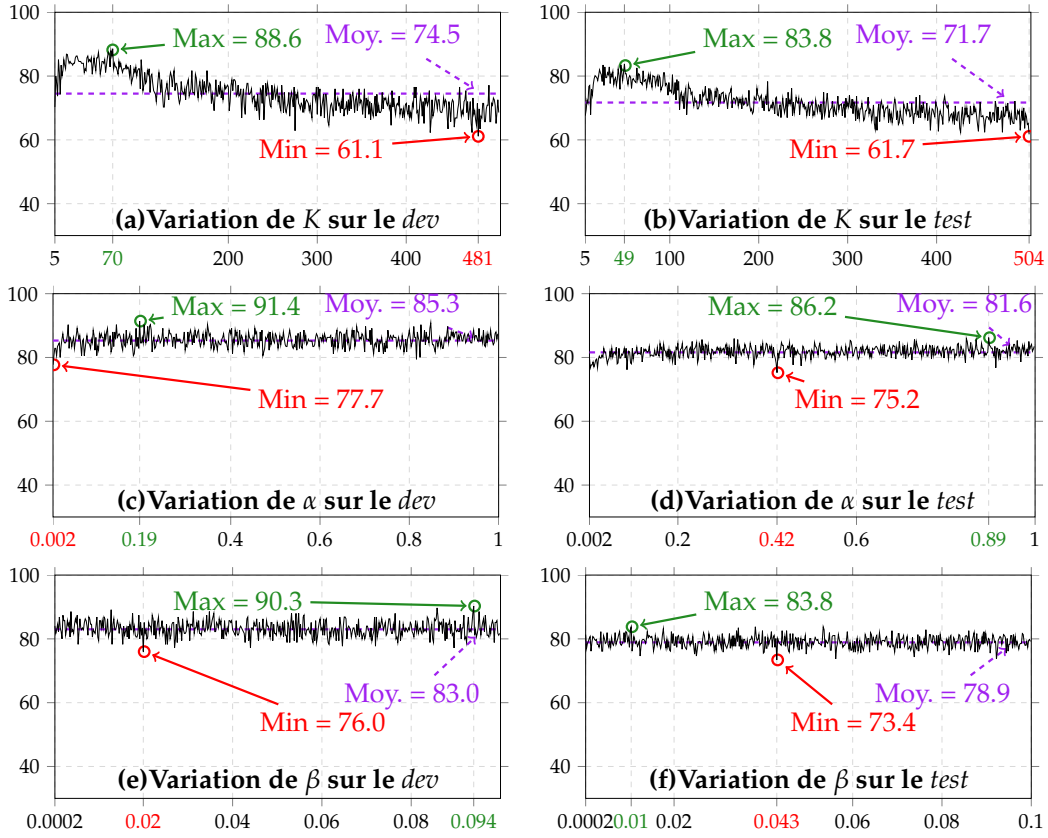


FIGURE 4.5 – *Précision de classification (%) en utilisant un ensemble de représentations basées sur les espaces de thèmes avec la normalisation EFR pour les données issues de l'ensemble de Dev et de Test du corpus Decoda. Différentes configurations expérimentales pour les ensemble d'apprentissage et de validation (Dev/Test) sont évaluées(TMAN). L'axe des abscisses représente la variation du le nombre K de thèmes contenus dans l'espace de thèmes ((a)-(b)) ; de α ((c)-(d)) et de β ((e)-(f)).*

(α, β et K). Les deux colonnes "Best" présentent les meilleures précisions obtenues avec le corpus de *Dev* et le corpus de *Test*, la dernière colonne présente les résultats obtenus quand la meilleure configuration trouvée avec l'ensemble de documents issus du *Dev* est appliquée à l'ensemble de *Test* (configuration réelle sachant que dans un cas réel de catégorisation de documents, l'étiquette associée à chacun des documents n'est pas connue).

Nous pouvons dans un premier temps noter que cette représentation compacte permet d'améliorer fortement les résultats obtenus par le système *baseline*, avec un gain de 8 points pour l'ensemble de *Dev* et 6,4 (86,5 (en variant K) $-80,1$(en variant α) $= 6,4$) points pour l'ensemble de *Test* en condition *Best* et $86,5 - 76,1 = 10,4$ points dans le cas réel pour le TRAP par exemple.

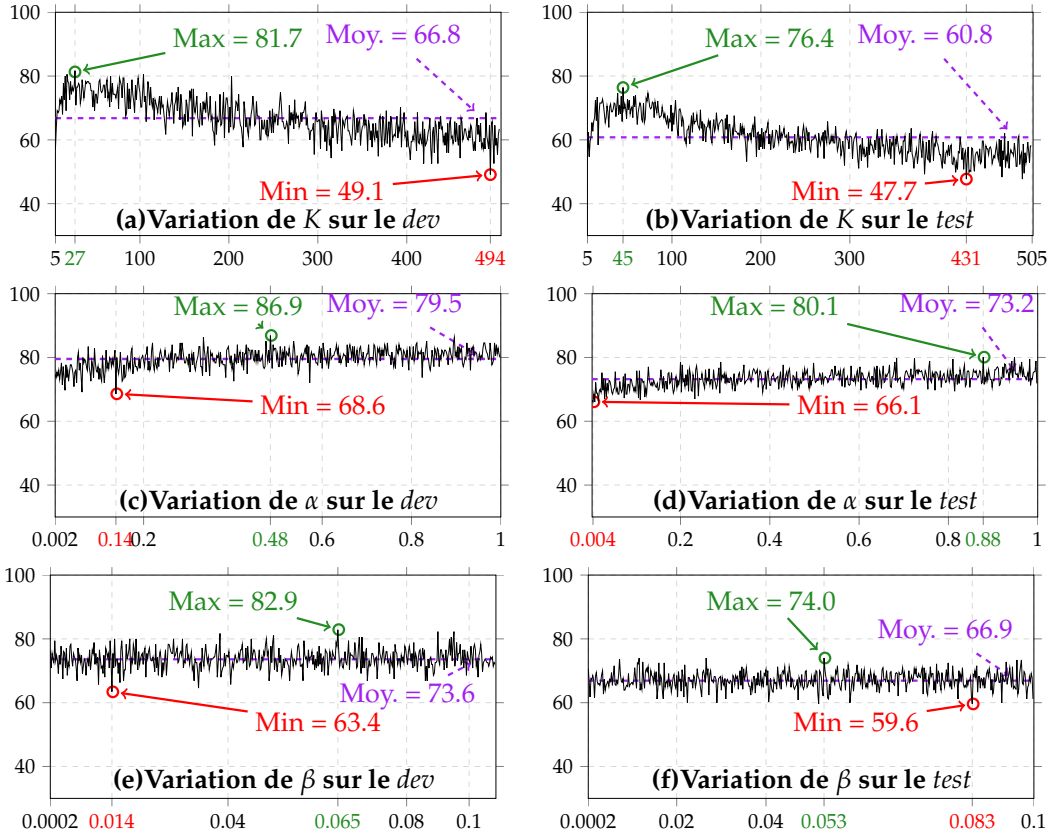


FIGURE 4.6 – *Précision de classification (%) en utilisant un ensemble de représentations basées sur les espaces de thèmes avec la normalisation EFR pour les données issues de l'ensemble de Dev et de Test du corpus Decoda. Différentes configurations expérimentales pour les ensemble d'apprentissage et de validation (Dev/Test) sont évaluées (TRAP). L'axe des abscisses représente la variation du le nombre K de thèmes contenus dans l'espace de thèmes ((a)-(b)) ; de α ((c)-(d)) et de β ((e)-(f)).*

4.4.6.2 Représentation compacte de documents textuels

La figure 4.7 montre la macro-F1 obtenue lors de la tâche de catégorisation d'articles issus du corpus Reuters-21578 en faisant varier différents hyper-paramètres du modèle LDA (n, α and β). La première remarque est que les meilleures précisions sont obtenues lorsque α varie, comme cela a déjà été observé pour le corpus DECODA (91,1 % pour le *Dev* et 76,8 % pour le *Test*), ensuite lorsque l'hyper-paramètre β varie (87,6 % pour le *Dev* et 74,1 % pour le *Test*) et enfin lorsque l'on varie le nombre de thèmes K (82,9 % pour le *Dev* et 72,6 % pour le *Test*).

Même si la différence Δ de la macro-F1, tous paramètres confondus, est significative pour le *Dev* d'une valeur de paramètre à l'autre ($3,5 \leq \Delta \leq 8,2$), cette différence se réduit quand le *Test* est utilisé ($1,5 \leq \Delta \leq 4,2$).

Le tableau 4.6 présente la macro-F1 obtenue avec la représentation compacte proposée couplée avec l'algorithme de standardisation EFR et différentes tailles de *c*-vecteurs

TABLE 4.5 – *Meilleures précisions et précisions réelles (%) obtenues lors de la tâche de catégorisation de conversations DECODA avec différentes méthodes et différentes configurations.*

Méthode employée	Variation du paramètre	Données		Best DEV	Best TEST	Réelle TEST
		TRAIN	TEST			
TbT	K	TMAN	TMAN	88.6	83.8	81.6
c-vecteur	K	TMAN	TMAN	92.0	**91.7**	**89.3**
TbT	K	TRAP	TRAP	81.7	76.4	70.6
c-vecteur	K	TRAP	TRAP	89.7	**86.5**	**86.5**
TbT	α	TMAN	TMAN	91.4	86.2	83.2
c-vecteur	α	TMAN	TMAN	88.3	**87.2**	**85.6**
TbT	α	TRAP	TRAP	86.9	80.1	76.1
c-vecteur	α	TRAP	TRAP	86.9	**83.2**	**80.4**
TbT	β	TMAN	TMAN	90.3	83.8	81.0
c-vecteur	β	TMAN	TMAN	87.4	**85.0**	**84.7**
TbT	β	TRAP	TRAP	82.9	74.0	68.5
c-vecteur	β	TRAP	TRAP	85.7	**80.4**	**75.4**

ainsi qu'un nombre variable de gaussiennes dans le GMM- UBM pour le corpus d'articles Reuters-21578. Les résultats présentés dans le tableau 4.7 montrent les scores obtenues en termes de macro-F1 en faisant varier différents paramètres du modèle LDA (α, β and n).

La première remarque est que les résultats obtenus avec la représentation compacte proposée (81,6 % et 80,0 % pour le *Test* dans les conditions respectives *Best* et réelles en faisant varier l'hyper-paramètre n), améliorent ceux obtenus avec la meilleure configuration des hyper-paramètres LDA (76,8 % et 70,9 % pour le *Test* dans les conditions respectives *Best* et réelle en faisant varier l'hyper-paramètre α) avec un gain de $4,8$ et $9,1$ points pour le *Test* dans les conditions respectives *Best* et réelle.

Nous pouvons également relever que les meilleurs résultats sont obtenus en faisant varier α pour le système *baseline* quand la meilleure macro-F1 est obtenue en faisant varier la granularité de l'espace de thèmes pour la version compacte c-vecteur proposée dans ce chapitre. Ces résultats confirment l'intuition initiale que l'hyper-paramètre α contrôle le processus d'inférence d'un nouveau document (n'apparaissant pas dans le corpus d'apprentissage) et que la représentation multi-vues nécessite un ensemble d'espaces de thèmes avec des granularités (*thmes*) différentes. Cette dernière approche est alors plus appropriée pour la représentation multiple d'un document compacté par le paradigme des c-vecteurs.

Cependant le tableau 4.7 montre que les macros-F1 obtenues avec le meilleur *Test* (pénultième colonne) sont les plus différentes pour la *baseline* TbT et c-vecteur, dans le contexte réel (meilleure configuration obtenue avec le *Dev* appliqué au *Test*). De plus, la différence Δ de la macro-F1 pour la représentation *baseline* TbT, quand α et β varient pour le *Test*, est faible ($\Delta = 2,5$ et $\Delta = 1,8$ points dans les conditions respectives *Best* et réelle). Cette différence est plus visible avec la version compacte où $\Delta = 4,1$ points pour le *Test* en faisant varier les hyper-paramètres α et β.

TABLE 4.6 – *Macro-F1 (%) avec différentes tailles de c-vecteurs et un nombre variable de gaussienne dans le GMM-UBM pour le corpus Reuters en faisant varier les hyper-paramètres α, β ainsi que le nombre de classes n contenues dans l'espace LDA.*

(a) Variation du nombre de thèmes K

taille du	*DEV*					*TEST*				
	Nombre de gaussiennes dans le GMM-UBM									
***c*-vecteur**	*32*	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	81.6	84.7	84.0	79.4	80.6	80.6	79.8	80.7	79.3	80.2
80	82.6	77.5	83.4	87.4	75.5	81.6	76.7	80.3	80.3	80.4
100	80.0	81.1	**86.7**	78.8	75.8	81.0	78.3	**80.0**	79.0	79.6
120	72.9	82.7	86.1	76.6	74.4	81.1	79.1	78.9	78.2	79.0
140	74.5	80.1	82.3	82.0	85.1	80.7	79.5	76.5	77.0	79.8
160	81.7	77.4	74.4	78.6	77.7	79.9	79.9	78.6	75.2	78.7
180	72.1	80.3	72.3	71.4	70.8	78.9	80.3	74.3	73.7	77.2
200	73.3	79.8	74.7	79.8	77.6	79.7	81.5	75.8	73.5	76.7

(b) Variation du paramètre α

taille du	*DEV*					*TEST*				
	Nombre de gaussiennes dans le GMM-UBM									
***c*-vecteur**	*32*	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	83.3	75.2	81.3	80.1	74.2	75.2	78.4	77.9	78.2	79.6
80	79.8	64.8	75.0	70.3	80.8	77.1	76.8	78.0	77.2	78.9
100	76.4	75.8	63.2	69.1	70.9	77.7	77.1	76.9	78.0	78.2
120	75.0	81.3	67.9	77.1	71.9	75.1	77.3	77.7	73.1	77.3
140	79.5	**86.2**	65.1	72.0	74.7	74.1	**74.5**	76.7	74.4	75.7
160	71.5	76.7	72.0	78.1	63.4	78.6	79.5	73.3	73.3	71.4
180	70.0	78.0	78.6	71.1	58.0	74.3	76.2	73.7	73.3	69.5
200	62.6	84.0	73.4	64.3	58.2	69.1	79.6	74.4	72.0	69.1

(c) Variation du paramètre β

taille du	*DEV*					*TEST*				
	Nombre de gaussiennes dans le GMM-UBM									
***c*-vecteur**	*32*	*64*	*128*	*256*	*512*	*32*	*64*	*128*	*256*	*512*
60	74.3	65.2	71.7	77.7	79.4	79.3	77.1	79.1	79.7	78.6
80	66.3	**80.8**	67.1	68.2	75.7	73.4	**78.6**	75.6	75.7	77.4
100	62.3	69.0	59.0	67.6	67.1	75.9	75.4	73.3	70.8	77.2
120	64.3	67.8	62.6	66.1	70.8	74.3	77.8	68.1	70.0	75.0
140	65.8	63.0	68.7	68.3	69.8	77.3	76.4	67.1	66.1	73.6
160	61.7	72.0	66.5	73.3	61.9	79.5	79.4	63.5	67.7	72.1
180	61.8	71.1	62.7	68.8	66.2	79.5	73.4	65.0	69.2	64.5
200	61.5	68.3	66.5	75.5	65.4	81.5	74.4	63.1	67.2	61.9

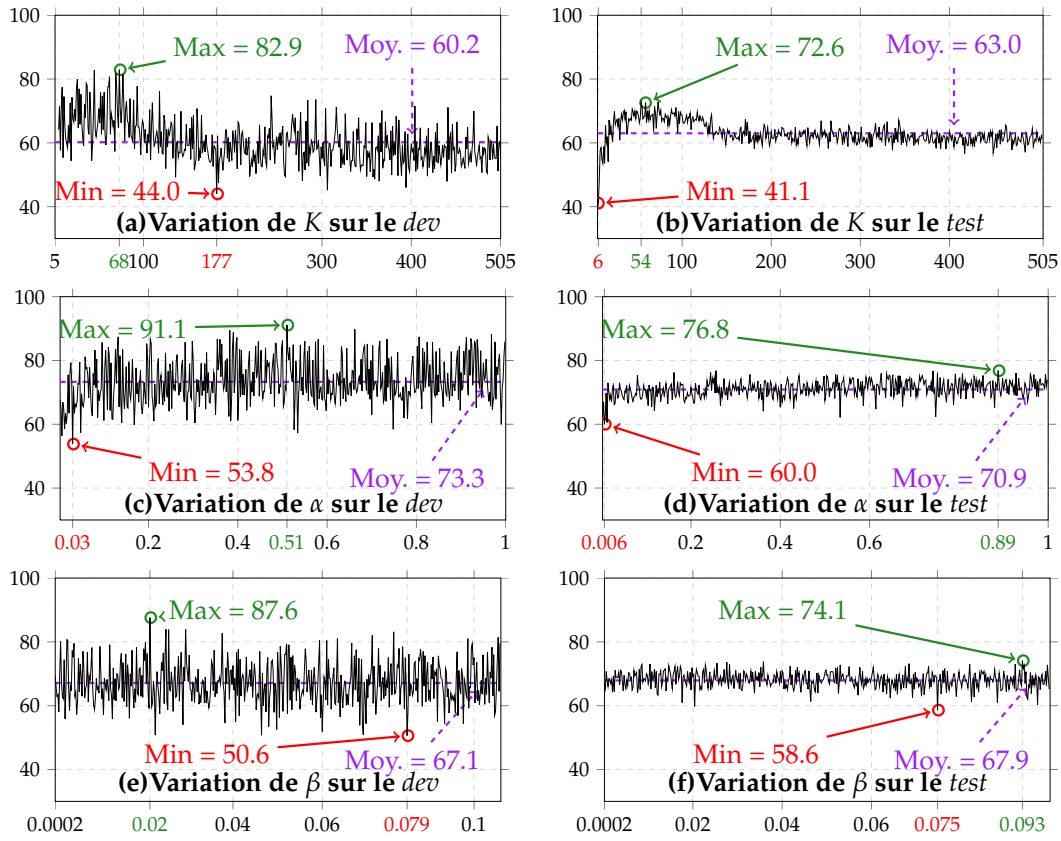


FIGURE 4.7 – *Macro-F1 (%) en utilisant un ensemble de représentations basées sur les espaces de thèmes avec la normalisation EFR pour les données issues de l'ensemble de Dev et de Test du corpus Reuters. L'axe des abscisses représente la variation du le nombre K de thèmes contenues dans l'espace thématique ((a)-(b)) ; de α ((c)-(d)) et de β ((e)-(f)).*

TABLE 4.7 – *Macro-F1 (%) obtenues dans des conditions Best et réelle lors de la tâche de catégorisation d'articles Reuters avec différentes méthodes et différentes configurations.*

Méthode employée	Variation du paramètre	Best DEV	Best TEST	Réelle TEST
TbT	K	82.9	72.6	63.6
c-vecteur	K	86.7	**81.6**	**80.0**
TbT	α	91.1	76.8	70.9
c-vecteur	α	86.2	**79.6**	**74.5**
TbT	β	87.6	74.1	69.1
c-vecteur	β	80.8	**79.5**	**78.6**

Nous pouvons conclure que cette approche originale, fondée sur les c-vecteurs, permet de traiter la variabilité contenue dans une conversation enregistrée dans un milieu bruité, comportant des segments (mots) traitant de thèmes différents selon le locuteur. Dans le contexte de la catégorisation automatique, cette technique améliore la préci-

sion, les résultats étant plus consistants quand la taille du *c*-vecteur ou le nombre de gaussiennes varient.

4.5 Conclusion sur l'apport des méthodes issues de l'analyse factorielle pour une représentation robuste de documents bruités

Le chapitre 3 a présenté l'apport d'une représentation dans de multiples espaces de thèmes LDA d'un document fortement bruité, lors d'une tâche de catégorisation ou de classification d'articles Reuters. Une des faiblesses des représentations dans des espaces thématiques par LDA est le choix des paramètres des modèles. Nous avons proposé une méthode qui consiste à estimer plusieurs espaces de thèmes correspondant à des configurations différentes. Les documents sont projetés dans ces espaces pour obtenir plusieurs "vues" du même document. Cette représentation multiple introduit de la redondance entre les vues et une variabilité potentiellement nuisible à la tâche de catégorisation. Nous avons proposé une méthode basée sur l'analyse factorielle pour extraire l'information utile de ces vues multiples, supprimer les variabilités nuisibles et "compacifier" l'espace de représentation. Cette méthode a été évaluée avec succès sur une tâche de catégorisation dans un corpus de parole et dans un corpus de document textuels.

La méthode proposée repose sur une première étape d'expansion de la représentation du document en vues complémentaires, suivie d'une phase de réduction de la représentation par extraction du sous-espace de variabilité utile. Ce principe pourrait être appliqué à divers problèmes de représentation multi-vues des données. Elle repose cependant sur une représentation des documents textuels en sac-de-mots, sans que l'ordre des mots ne soit pris en compte. L'influence de cette information peut être limitée lorsque la structure temporelle des documents est faible. Dans le cas général, et en particulier dans celui de dialogues très structurés, la position des termes est une caractéristique qui est essentielle. Le chapitre suivant introduit une représentation originale fondée sur les nombres hyper-complexes permettant d'introduire, dans le modèle de documents, des informations relatives à la structure temporelle des documents.

Troisième partie

Caractéristiques hyper-complexes de termes bruités

Chapitre 5

Projection de documents bruités dans l'espace hyper-complexe des Quaternions

Sommaire

Résumé

Les représentations dans des espaces thématiques permettent de modéliser une structure sémantique sous-jacente extraite de grands corpus ; cependant, elle repose sur une représentation en sacs de mots dans laquelle la structure temporelle des documents est perdue. Dans ce chapitre, nous proposons une représentation basée sur des nombres hypercomplexes qui permet de ré-introduire cette information temporelle tout en préservant les propriétés d'"estimabilité" du modèle sur des quantités de données limitées.

5.1 Introduction

Les performances des systèmes de reconnaissance automatique de la parole (SRAP) sont très dépendantes des conditions d'enregistrement. Lorsque ces conditions sont optimales, les transcriptions issues des SRAP sont de bonne qualité, ce qui de leurs appliquer des méthodes issues du traitement de document écrit classique. Cet environnement favorable n'est que très rarement rencontré dans des cas d'utilisations réelles.

Toutes ces représentations du document dans des espaces de thèmes différents, s'appuient sur LDA. Le modèle LDA considère le document comme un "sac-de-mots", sans tenir compte des relations entre les occurrences d'un même terme dans le document, hormis la distribution des termes au sein des thèmes du modèle. Cette relation n'est pas explicite, dans le sens où la variable latente liant les termes aux classes ne tient pas compte de la position de ces termes au sein du document.

L'importance de cette hypothèse prend tout son sens dans le cas de documents très structurés, par exemple des conversations entre client et agents d'un centre d'appels.

Plusieurs méthodes ou caractéristiques ont été proposées dans la littérature, pour tenir compte de la position relative d'un terme au sein du document. Parmi ces indicateurs, il y a la position dite "relative", ou première occurrence du terme dans le document (RP), décrite dans la section A.3.2. Cette caractéristique ne tient compte que de la position de la première occurrence du terme et ne tient pas compte des autres occurrences de ce terme ou de la distribution des positions occupées par ce mot dans le document.

Une représentation idéale devrait tenir compte de la position de chacune des occurrences du mot dans le document, ainsi que de la relation liant les occurrences de ce mot au sein de ce document. L'information classique de la fréquence du terme est codée dans un nombre réel compris entre 0 et 1. La représentation étudiée dans ce chapitre considère la position du terme dans le document selon que celui-ci se trouve en début, au milieu, ou en fin de document. Pour coder cette information multiple dont les composantes sont liées, un nombre hyper-complexe est utilisé. Ce nombre contient 4 "parties" pouvant alors contenir 4 caractéristiques du terme dans le document. Il est issu de l'algèbre de Clifford et est appelé *quaternion*.

Un étude mesurant l'apport d'une telle représentation de documents bruités dans l'espace des hyper-complexes est ainsi proposée. Cet espace tient compte de la position du terme au sein du document et des relations entre les différentes occurrences du terme dans le document. Cette relation est prise en compte lors des différentes opérations mathématiques entre les quaternions.

Le chapitre est organisé comme suit. Dans un premier temps, l'algèbre des quaternions est présentée dans la section 5.2. La section 5.3 présente les différents domaines d'application de l'algèbre des quaternions. Une étude portant sur la catégorisation de documents bruités projetés dans l'espace des hyper-complexes, est détaillée dans la section 5.4. Une conclusion sur l'apport d'une telle représentation est donnée finalement dans la section 5.5 pour des documents fortement bruités.

5.2 Les Quaternions

Les quaternions sont des nombres hyper-complexes, leur ensemble, noté $\mathbb{H}$, pouvant être vus comme une extension de l'ensemble des complexes $\mathbb{C}$ à $\mathbb{R}^4$. L'origine des quaternions date du $XIX^{\text{ème}}$ siècle, quand William Rowan Hamilton (Hamilton, 1866) découvrit les quaternions en 1843. Son objectif d'alors était de construire une algèbre avec des triplets de nombres réels. Lors de ses recherches, il butait sur la multiplication et la conservation des normes pourtant considérée comme impossible par Frobenius (Frobenius, 1877). Hamilton proposa alors d'utiliser une quatrième composante et introduit ce nouveau quadruplet, appelé *quaternion*, comme une nouvelle entité des nombres complexes.

Un quaternion Q est un quadruplet $\{r, x, y, z\}$ de nombres réels composé d'un nombre scalaire r et d'un vecteur $\vec{v}$ de trois composantes $xi + yj + zk$ considéré comme la partie imaginaire pure du quadruplet :

$$Q = r1 + xi + yj + zk \tag{5.1}$$
$$= r + xi + yj + zk \tag{5.2}$$
$$= r + \vec{v} \tag{5.3}$$

Cette algèbre n'est pas commutative mais partiellement anti-commutative. En effet, les composantes du quaternion respectent les règles de multiplication définies dans le tableau 5.1 : si l'on a bien $1i = i1$, en revanche nous avons le produit $ij = -ji = k$.

$\times$	**1**	**i**	**j**	**k**
1	1	i	j	k
i	i	-1	k	$-j$
j	j	$-k$	-1	i
k	k	j	$-i$	-1

TABLE 5.1 – *Table de multiplication entres composantes du quaternion.*

Les quaternions répondent à un ensemble de propriétés :

- Q est nul si : $r = x = y = z = 0$,
- les ensembles des quaternions de la forme $Q = r1 + 0i + 0j + 0k = r$ sont appelés scalaires ou réels,
- comme l'indique la diagonale du tableau 5.1 :

$$i^2 = j^2 = k^2 = ijk = -1 \,, \tag{5.4}$$

- le produit formel noté $**$ entre deux quaternions $Q = r1 + xi + yj + zk$ et $Q' = r'1 + x'i + y'j + z'k$, et appelé multiplication d'Hamilton, se fait en développant le produit terme par terme, puis en appliquant les règles de réduction définies dans l'équation 5.4 :

$$Q * *Q' = (rr' - \langle \vec{v}, \vec{v}' \rangle) + (r\vec{v} + r'\vec{v} + \vec{v} \wedge \vec{v}') \,, \tag{5.5}$$

avec,

$$\vec{v} = xi + yj + zk \text{ et } \vec{v}' = x'i + y'j + z'k \tag{5.6}$$

où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire et $\wedge$ le produit vectoriel entre les deux parties vectorielles des deux quaternions. La multiplication entre deux quaternions n'est pas commutative et il existe Q et Q' tels que :

$$Q * *Q' \neq Q' * *Q \,, \tag{5.7}$$

- le conjugué $\overline{Q}$ du quaternion Q est :

$$\begin{aligned} \overline{Q} &= r1 - xi - yj - zk && (5.8)\\ &= r1 - (xi + yj + zk) && (5.9)\\ &= r1 - \vec{v} && (5.10) \end{aligned}$$

- la norme d'un quaternion est :

$$\begin{aligned} |Q| &= \sqrt{Q * *\overline{Q}} \\ &= \sqrt{\overline{Q} * *Q} \\ &= \sqrt{r^2 + x^2 + y^2 + z^2} \end{aligned} \tag{5.11}$$

- l'inverse unique du quaternion Q est désigné par Q^{-1} :

$$Q^{-1} = \frac{\overline{Q}}{|Q|^2} \tag{5.12}$$

vérifie :

$$Q * *Q^{-1} = Q^{-1} * *Q = 1 \tag{5.13}$$

- un quaternion unitaire (normalisé) $Q^{\triangleleft}$ est défini comme suit :

$$Q^{\triangleleft} = \frac{Q}{|Q|^2} \tag{5.14}$$

- enfin, le produit scalaire entre deux quaternions Q et Q' est donné par :

$$\langle Q, Q' \rangle = rr' + xx' + yy' + zz' \tag{5.15}$$

Plus d'informations concernant les systèmes fondés sur l'algèbre des hypercomplexes sont disponibles dans (Kantor et al., 1989; Zhang, 1997; Kuipers, 1999), et plus précisément sur les quaternions dans (Ward, 1997).

5.3 Domaines d'application des quaternions

La non-commutativité des quaternions peut être relativement intuitive si l'on prend l'exemple d'un objet dans un espace de dimension trois. En appliquant une rotation autour de l'axe $\vec{X}$ puis une rotation autour de l'axe $\vec{Y}$, l'objet n'est pas dans la même position que si on lui applique les deux rotations dans l'ordre inverse. Ceci est le fondement de l'ensemble des quaternions. En effet, si nous nous reportons au domaine de la géométrie, la représentation à l'aide des quaternions permet de distinguer deux rotations d'une même similitude, en introduisant deux points intermédiaires permettant de distinguer les deux trajets. Il est donc naturel que les chercheurs du traitement de l'image (voir section 5.3.1) et de la modélisation des mouvements (voir section 5.3.2) soient les premiers à employer les propriétés de rotation de l'algèbre des quaternions dans leurs travaux de recherche. Plus généralement (voir section 5.3.3), les utilisateurs de méthodes de traitement de l'information trouvent également, dans l'algèbre des quaternions, une représentation robuste et élégante permettant de structurer l'information.

5.3.1 Traitement d'images à l'aide des quaternions

L'utilisation des quaternions dans le domaine du traitement d'images date de plus de vingt ans maintenant. Les méthodes proposées représentent les images en couleurs en codant trois informations dans les parties imaginaires ($\vec{v}$) des quaternions (Pei et Cheng, 1999; Le Bihan et Sangwine, 2003a; Pei et al., 2003; Le Bihan et Sangwine, 2003b; Alexiadis et Sergiadis, 2009; Assefa et al., 2010; Chen et al., 2010; Guo et Zhu, 2011; Sun et al., 2011; Rizo-Rodríguez et al., 2013). Ces trois caractéristiques correspondent aux quantités de rouge (R), de vert (G), et de bleu (B) dans le codage RGB, et sont contenues respectivement dans les valeurs du quaternion x, y et z. L'avantage principal de la représentation dans des quaternions est qu'une image en couleur peut être traitée de manière globale comme un vecteur unique (Subakan et Vemuri, 2011; Ell et Sangwine, 2007). L'algèbre des quaternions a été exploitée dans le traitement d'images digitales par (Sangwine, 1996) ainsi que par (Pei et Cheng, 1997). Depuis, plusieurs outils, utilisés pour les niveaux de gris, ont été étendus en utilisant l'algèbre des quaternions, pour le traitement de l'image, avec un certain succès. Ces outils incluent les transformées de Fourier (Ell, 1992; Bülow, 1999), de Fourier-Melin (Hitzer, 2013), d'ondelettes (Chan et al., 2004; Bayro-Corrochano, 2006), de l'harmonique polaire (Li, 2013), et leurs moments (Chen et al., 2010; Guo et Zhu, 2011; Chen et al., 2012).

Récemment, l'algèbre de Clifford (Clifford, 1873), aussi connue comme une algèbre géométrique généralisant l'algèbre des quaternions, a permis le traitement de signaux de plus grande dimension, comme le sont les images en trois dimensions. Les processus permettant ce traitement sont, par exemple, les transformées de Clifford Fourier (Batard et al., 2010; Mennesson et al., 2011; Hitzer et Sangwine, 2013), les moments de Clifford Fourier-Melin (Mennesson et al., 2014), les ondelettes de Clifford (Bahri et Hitzer, 2007), ainsi que la corrélation géométrique (Bujack et al., 2014).

5.3.2 Les quaternions dans la gestion de mouvements

Comme il est vu dans la section 5.3.1, les quaternions ont également été utilisés dans le domaine du traitement d'images couleurs en trois dimensions (3D). Le cas réel le plus approchant des images 3D, est la gestion de mouvements dans l'espace. (Shuster, 1993; Xian et al., 2004; Yuan, 1988) et (Daniilidis, 1999; Wu et al., 2005) montrent respectivement que lors de l'enregistrement de mouvement spatiaux, l'orientation d'un corps peut être paramétrisée en utilisant un quaternion unitaire, comme défini dans l'équation 5.14, ainsi que des *dual-quaternions*.

Le quaternion unitaire est une représentation en quatre paramètres contenant une seule contrainte. Par conséquent, il fournit la plus petite dimension possible pour une représentation globale et sans singularité de l'attitude (ou orientation). Bien que ce quaternion unitaire représente l'orientation de l'extrémité du bras du robot sans singularité, la réalisation du régulateur devient compliquée sachant que les erreurs d'orientation sont calculées séparément (Xian et al., 2004; Yuan, 1988). En effet, les schémas de contrôle de la position sont habituellement réalisés en deux boucles de contrôle, pour contrôler dans le même temps la rotation et la translation (Sciavicco et Villani, 2009).

En dépit du nombre de dimensions du dual-quaternion (8), (Aspragathos et Dimitros, 1998; Funda et al., 1990) montrent que cette représentation est la manière la plus compacte et la plus performante de caractériser les mouvements qui sont, dans un même temps, des transformations transversales et rotationnelles dans la chaîne kinématique d'un robot. Le dual-quaternion est un outil élégant très utile pour l'analyse kinématique dans bien des recherches, comme la navigation inertielle (Wu et al., 2005) ainsi que dans la vision par ordinateur (Daniilidis, 1999).

5.3.3 Méthodes génériques bâties sur l'algèbre des quaternions

La plupart des tâches de catégorisation du domaine du traitement du langage naturel nécessitent l'utilisation d'une méthode permettant d'associer, à un document donné, une étiquette. Les plus utilisées sont les méthodes fondées sur les réseaux de neurones ainsi que les machines à vecteurs de support (SVM). Ainsi, ces méthodes ont été adaptées avec un certain succès dans l'espace hyper-complexe des quaternions par (Arena et al., 1994; Nitta, 1995; Pearson et Bisset, 1994) pour respectivement les réseaux de neurones de quaternions et dans l'algèbre de Clifford, et par (Bayro-Corrochano et Arana-Daniel, 2010) pour les SVM de Clifford.

La tâche de réduction des espaces de représentation a également permis d'adapter les outils classiques en utilisant l'algèbre des quaternions comme la décomposition en valeurs singulières (Le Bihan et Sangwine, 2003a; Pei et al., 2003), l'analyse en composantes principales (Le Bihan et Sangwine, 2003b; Guo et Zhu, 2011), et l'analyse en composantes indépendantes (Le Bihan et al., 2006; Vía et al., 2011).

Toutes ces applications de l'algèbre des quaternions à des domaines très variés, montrent qu'une représentation de données dans un espace hyper-complexe, permet

de mieux les structurer et de formaliser la relation statistique entre les composantes de ce nombre. La section 5.4 qui suit étudie l'apport d'une telle représentation dans l'objectif de formaliser les relations existantes entre les occurrences d'un même mot au sein d'un document. Ainsi, chacun des termes composant le document ne sera plus considéré comme indépendant de sa position dans le document, comme cela est le cas lors de l'utilisation d'une représentation dans un espace de thèmes LDA.

5.4 Représentation de documents bruités par des quaternions

Dans le cas de conversations entre un utilisateur et un agent, ce dernier doit détecter au plus tôt le sujet de l'appel. Ainsi, l'information permettant à l'agent de déterminer la raison de l'appel de l'usager est contenue essentiellement dans des segments se trouvant à des positions variables selon le motif de l'appel. Ainsi, une personne désirant faire une réclamation pour un procès verbal indûment reçu, dû à un défaut de la carte de transport, aura tendance, dans un premier temps, à parler de sa carte de transport et des différentes zones que celle-ci recouvre, avant de parler de l'infraction elle-même. Pour cette raison, la position des termes tient un rôle important pour localiser, de la manière la plus pertinente possible, l'information utile permettant de déterminer le motif exact lié à une conversation.

De plus, les dialogues sont redondants et guidés par l'agent de la Régie autonome des transports parisiens (RATP) qui suit, autant que possible, un schéma de dialogue standard. Considérant que cette structure de dialogue sous-jacente pourrait apporter une information capitale sur le thème de la conversation, il est proposé dans cette étude un modèle de conversation s'appuyant sur le découpage de ce dialogue en quatre parties de même taille ainsi que l'utilisation de l'algèbre des quaternions.

La méthode proposée consiste à étendre le modèle vectoriel classique représentant le document par un vecteur de fréquences de termes (Salton, 1989b) : ici, un document est représenté par un vecteur de quaternions dans lequel chacune des dimensions est un quaternion regroupant les quatre fréquences d'un terme dans chacune des quatre parties de la conversation. L'intuition centrale est que ce modèle s'appuyant sur les quaternions est plus pertinent qu'une représentation classique utilisant les fréquences de mots sur l'ensemble du document. Cette représentation permet, en effet, de combiner l'impact de différentes phases du discourt lors de l'élaboration de la représentation vectorielle. Cette méthode est developpée et évalué dans le cadre de l'analyse des dialogues du corpus DECODA (Bechet et al., 2012).

La section 5.4.1 présente la représentation d'un document utilisant un vecteur de quaternions. Les expérimentations portant sur une tâche de catégorisation de conversations utilisant les représentations classiques ainsi que la représentation fondée sur un vecteur de quaternions, sont présentées dans la section 5.4.2.

5.4.1 Système bâti sur une représentation vectorielle de quaternions

Une représentation classique d'un document ne permet pas de prendre en compte la position d'un terme dans un document. Ceci peut être une faiblesse lorsque l'on traite des documents pouvant contenir plusieurs sujets et ainsi, contenir un thème principal localisé à un ou plusieurs endroits précis dans le document. Pour pallier cette difficulté, la méthode proposée s'appuie sur la représentation d'un document par un vecteur de quaternions. Chacun de ces quaternions est la représentation d'un terme dans le document. Les quatre composantes de ce quaternion correspondent à la fréquence du terme à un instant précis de la conversation. Ces segments sont de tailles égales.

La section 5.4.1.1 présente le modèle classique fondé sur la fréquence de mots discriminants. La segmentation du dialogue en quatre parties de tailles égales est présentée dans la section 5.4.1.2. La section 5.4.1.3 est dévolue à la mesure de similarité entre deux vecteurs de quaternions. Enfin, la section 5.4.1.4 présente la méthode de catégorisation de dialogues bâtie sur une représentation vectorielle de quaternions.

5.4.1.1 Extraction d'un ensemble de termes discriminants

Dans le but d'estimer les paramètres de différents espaces de thèmes [1] et d'obtenir une représentation commune [2], un ensemble de mots discriminants V est constitué comme décrit dans la section 2.2.3.1. Les $|V|$ termes w obtenant le score $\Delta(w)$ le plus élevé pour chacun des catégories t, sont conservés pour constituer le vocabulaire V. Ainsi, chacune des catégories t (parmi les T catégories) est associé à une liste de termes V_t le décrivant au mieux. Notons, néanmoins, qu'un même mot peut apparaître dans plusieurs listes de termes discriminants V_t. Tous ces termes sont ensuite regroupés, sans répétition, dans une liste de termes discriminants V :

$$V = \bigcup_{t=1}^{T} V_t \tag{5.16}$$

5.4.1.2 Segmentation du dialogue

La méthode proposée est motivée par l'hypothèse que la structure d'un dialogue offre un point de vue précis de la structure thématique permettant d'améliorer les performances du système de catégorisation de dialogue.

Un des points clés, pour la capture de caractéristiques structurales, est la segmentation. Idéalement, la segmentation devrait correspondre aux différentes phases de la

1. $P(z|d)$ avec différents z pour chacun des espaces de thèmes.
2. Chacune des caractéristiques du vecteur de représentation du document doit correspondre à une caractéristique identique d'un espace de thèmes à un autre.

conversation, mais en pratique, ces phases sont de tailles variables, pouvant se juxtaposer. Ceci ne permet pas de dessiner des frontières claires du dialogue. Ici, la segmentation est utilisée comme un moyen d'extraire des caractéristiques permettant de suivre les variations de thèmes tout au long de la conversation.

Deux schémas de segmentation sont testés (voir figure 5.1). Le premier schéma consiste en une segmentation "naturelle" de gauche à droite (LRQ) : le document est, dans un premier temps, partagé en cinq régions, chacune représentant 20 % du document. Ensuite, un segment est constitué de deux régions successives de 20 % pour finalement couvrir 40 % du document.

Le second schéma est une segmentation symétrique (SSQ) s'appuyant sur la position du mot relativement au centre du dialogue. Dans ce schéma de segmentation, les quatre régions correspondent à :

- (1) la première moitié du document,
- (2) la région centrale (50 % centré),
- (3) la seconde moitié du document,
- (4) la jonction entre les deux extrémités du document (premier et dernier 25 %).

La figure 5.1 montre un exemple d'une segmentation d'un document avec le schéma de gauche à droite (S_1, S_2, S_3, S_4) et le schéma symétrique (W_1, W_2, W_3, W_4).

Ces deux stratégies de segmentation permettent d'extraire quatre statistiques d'un terme qui seront encapsulées dans un quaternion, comme le décrit la section suivante.

5.4.1.3 Représentation d'un document dans un vecteur de quaternions

Sachant une segmentation $S = \{s_1, s_2, s_3, s_4\}$ d'un document $d \in D$ ainsi qu'un vocabulaire de termes discriminants $V = \{w_1, \ldots, w_n, \ldots, w_{|v|}\}$, chaque terme w_n est représenté par le quaternion :

$$Q_d(w_n) = f_d^1(w_n)1 + f_d^2(w_n)i + f_d^3(w_n)j + f_d^4(w_n)k \tag{5.17}$$

où $f_d^m(w_n)$ représente la fréquence du mot w_n dans le segment s_m du document d.

La figure 5.1 illustre la représentation fondée sur les quaternions d'un document avec les deux schémas de segmentation, et montre un dialogue réel issu du corpus de validation étiqueté par l'agent comme traitant d'un *problème d'itinéraire*. Un document $d \in D$ est représenté par un vecteur de quaternions Q_d de dimension $|V|$:

$$Q_d = [Q_d(w_n)]_{w_n \in V} \ . \tag{5.18}$$

Soient deux quaternions représentés par leurs parties réelles et imaginaires comme $Q_1 = r_1 + \vec{v_1}$ et $Q_2 = r_2 + \vec{v_2}$. Si un quaternion est normalisé comme indiqué dans l'équation 5.14, alors la somme des carrés des valeurs réelles (r, x, y et z) vaut 1. Ainsi, le quaternion représente une orientation et la distance entre deux quaternions de ce type

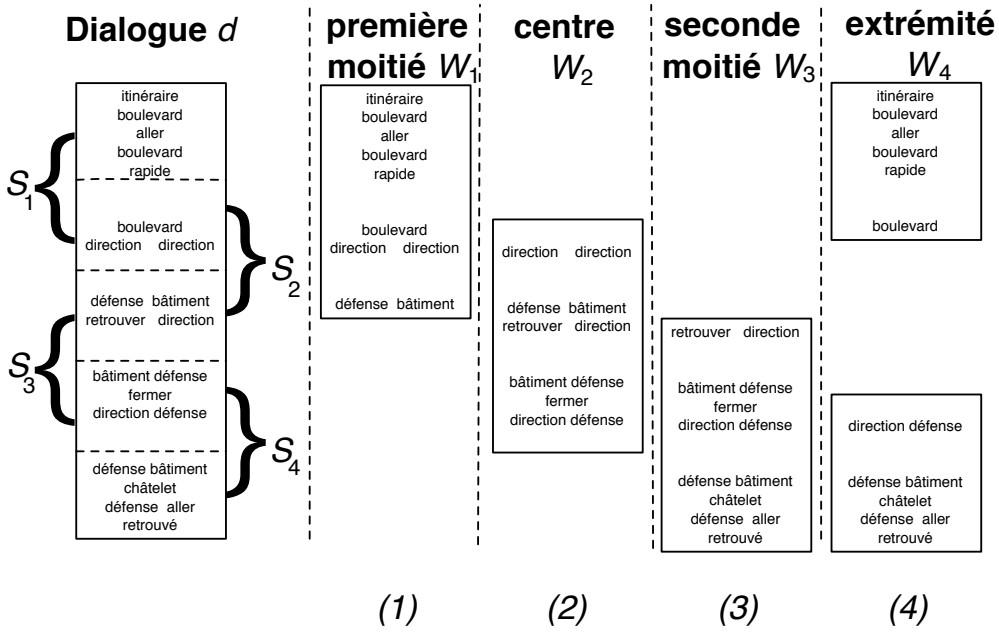


FIGURE 5.1 – *Dialogue provenant du corpus de validation segmenté selon le schéma de gauche à droite LRQ (S_1, S_2, S_3, S_4) et le schéma symétrique SSQ (W_1, W_2, W_3, W_4). Ce dialogue est correctement étiqueté par les deux méthodes fondées sur des vecteurs de quaternions, alors que la méthode classique fondée sur la fréquence des termes TF-IDF a échoué.*

correspond à la distance angulaire entre les deux orientations représentées par chacun des quaternions (axe de rotation). Définissons deux quaternions à l'aide de l'équation 5.17 représentant les distributions du même mot w dans deux documents distincts. Soient $Q^{\triangleleft}_{d_1}(w)$ et $Q^{\triangleleft}_{d_2}(w)$ les versions normalisées selon l'équation 5.14 des deux quaternions $Q_{d_1}(w)$ et $Q_{d_2}(w)$. La distance entre les deux documents résulte simultanément de la formulation de l'angle double du cosinus :

$$cos(2\phi) = 2cos^2(\phi) - 1 \tag{5.19}$$

et du fait que l'angle entre les orientations θ est précisément le double de l'angle ϕ entre les deux quaternions :

$$\phi = \frac{\theta}{2} \tag{5.20}$$

et donc,

$$\begin{aligned} cos(2\frac{\theta_{1,2}(w)}{2}) &= 2cos^2(\frac{\theta_{1,2}(w)}{2}) - 1 \\ &= 2\left\langle Q^{\triangleleft}_{d_1}(w), Q^{\triangleleft}_{d_2}(w)\right\rangle^2 - 1 \end{aligned}$$

ainsi ,

$$\theta_{1,2}(w) = cos^{-1}\left[2\left\langle Q^{\triangleleft}_{d_1}(w), Q^{\triangleleft}_{d_2}(w)\right\rangle^2 - 1\right]$$

où $\langle \cdot, \cdot \rangle$ est le produit scalaire défini dans l'équation 5.15. Un vecteur est obtenu en calculant la distance angulaire entre les quaternions représentant chacun des termes w pour les deux documents. Ensuite, la dissimilarité entre les deux conversations est le score moyen entre tous les termes discriminants contenus dans V :

$$\Theta(d_1, d_2) = \frac{1}{|V|} \sum_{w \in V} \theta_{1,2}(w) \tag{5.21}$$

5.4.1.4 Méthode de catégorisation

La méthode proposée, s'appuyant sur des vecteurs de quaternions, est plutôt un paradigme de représentation qu'une méthode de catégorisation. Plusieurs méthodes peuvent être appliquées pour prendre la décision du thème à associer au dialogue dans l'espace de caractéristiques des quaternions (Harish et al., 2010). Pour cette raison, des expérimentations contrastives sont conduites en utilisant la méthode des k plus proches voisins (la méthode des k plus proches voisins ou *K-Nearest Neighbor* (KNN)) (Bentley, 1975), qui requiert uniquement de définir une distance à appliquer à des caractéristiques non standards.

Dans la tâche d'identification de catégorie, l'algorithme des KNN calcule la distorsion entre une conversation issue du corpus de validation C_i et chacune des conversations contenues dans le corpus d'apprentissage. Un sous-ensemble $SS_k(C_i)$ de k plus proches éléments sont extraits. La probabilité d'une catégorie t sachant C_i est estimée en comptant le nombre de conversations de ce thème t dans SS_k. Dans nos expérimentations, le nombre de plus proches voisins k est fixé empiriquement selon le corpus de développement.

5.4.2 Expérimentations

Cette section décrit le protocole expérimental, ainsi que les résultats obtenus avec une représentation des dialogues utilisant des caractéristiques classiques basées sur la fréquence de mots (TF-IDF) pour chacun des segments puis sur l'ensemble du document. Ces résultats sont comparés avec ceux obtenus en utilisant des schémas de segmentation avec les quaternions.

Pour mesurer les performances d'une représentation tenant compte de la position des termes au sein du document, l'ensemble des conversations issues du corpus DECODA (Bechet et al., 2012) est utilisé dans une tâche de catégorisation. Ce corpus est détaillé dans la section 2.2.5 ainsi que le système de reconnaissance automatique de la parole (SRAP) utilisé pour transcrire les conversations entre agents et utilisateurs.

Données		*Précision (%)*									
Test	**Appr**	TS_1		TS_2		TS_3		TS_4		**M4D**	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
TMAN	TMAN	*70,3*	64,4	*69,1*	62,5	*68,0*	62,2	*61,1*	39,8	*86,2*	78,8
TRAP	TMAN	*65,1*	55,0	*66,8*	54,1	*61,7*	48,3	*53,7*	39,1	*81,7*	70,0
TRAP	TRAP	*64,0*	56,2	*65,7*	50,7	*63,4*	48,3	*53,1*	44,3	*78,1*	67,4
TRAP	TMAN+TRAP	*64,5*	57,1	*66,2*	50,7	*61,7*	50,1	*55,4*	45,5	*82,5*	70,3

TABLE 5.2 – *Résultats obtenus par les systèmes de base en termes de précision (%))*

5.4.2.1 Systèmes de base

Une comparaison en deux étapes est réalisée : la première compare six représentations fondées sur la fréquence de termes puis entre les deux représentations bâties sur les quaternions et la meilleure représentation fondée sur la fréquence de termes. Toutes ces configurations sont comparées en utilisant le même protocole expérimental, incluant un ensemble de mots discriminants communs, une méthode de catégorisation identique, ainsi que des corpus d'apprentissage, de développement, et de test communs.

Les deux systèmes évalués, s'appuyant sur les quaternions, utilisent les schémas de segmentation gauche à droite et symétrique, comme il est décrit dans la section 5.4.1. Ces schémas sont notés respectivement LRQ et SSQ.

Le premier système *baseline* s'appuie sur l'approche classique TF-IDF, dans lequel chacun des documents est caractérisé par un vecteur de fréquences de mots. Ces fréquences sont estimées pour toute la conversation, et la mesure de cosinus est utilisée comme une distance entre deux documents.

Un aspect important de la méthode proposée ici est le processus de segmentation fondé sur l'intuition que la distribution dépendante de la position du terme, permet de révéler un aspect important sur la distribution des thèmes durant la conversation. Le système proposé combine des caractéristiques dépendantes des phrases du dialogue (ou segments).

Dans l'objectif d'évaluer les gains obtenus par ces deux aspects séparément, un système s'appuyant sur le TF-IDF opérant sur des vecteurs de fréquences estimés sur les quatre segments successifs plutôt que sur l'ensemble du dialogue, est également testé dans un premier temps. Dans ce système appelé *M4D* par la suite, un vecteur représente un document incluant $4 * |V|$ coefficients (sac-de-mot de $|V|$ termes). La mesure de distance *cosinus* est utilisée lors de la phase de classification.

Cette méthode est directement comparable à l'approche bâtie sur les quaternions qui utilisent le même ensemble de caractéristiques de bas niveau (LRQ et SSQ). Ces caractéristiques sont composées de la fréquence des termes dans chacun des segments. Ceci permet d'évaluer l'intérêt spécifique des quaternions, indépendamment des caractéristiques basiques résultant de la segmentation.

La dernière approche groupe toutes les fréquences dépendant des segments, dans

un vecteur de grande taille. Un dernier test est effectué, dans lequel la contribution de chacune des parties de la conversation dans un schéma gauche à droite est vérifiée. Ceci est réalisé en évaluant la performance des systèmes fondés sur le TF-IDF utilisant uniquement les distributions des termes calculées sur un seul segment. Ces systèmes sont appelés TS_1, TS_2, TS_3 and TS_4 pour les segments, du premier au quatrième.

Tous ces systèmes sont évalués avec un classifieur KNN ainsi qu'avec un corpus d'apprentissage composé de dialogues issus de transcriptions manuelles (TMAN), de dialogues transcrits automatiquement (TRAP), et enfin la combinaison des deux (TMAN+TRAP).

Données		*Précision (%)*							
Test	**Apprentissage**	**TF-IDF**		**M4D**		**LRQ Θ**		**SSQ Θ**	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
TMAN	TMAN	72,0	71,7	86,2	78,8	89,0	85,3	92,2	87,4
TRAP	TMAN	68,4	64,4	81,7	70,0	82,2	74,1	84,6	76,0
TRAP	TRAP	65,0	61,3	78,1	67,4	80,5	71,7	82,2	73,9
TRAP	TMAN+TRAP	67,5	62,2	82,5	70,3	81,0	73,3	83,4	**76,9**

TABLE 5.3 – *Résultats en termes de précision.*

5.4.2.2 Résultats et discussions

Les expérimentations ont été réalisées dans le contexte de la catégorisation de conversations issues du corpus DECODA. Les sacs-de-mots (*BoW*) ont été utilisés dans le même objectif que les expérimentations rapportées dans (Koço et al., 2012; Maza et al., 2011). Dans (Maza et al., 2011), des caractéristiques liées aux phrases sont ajoutées aux caractéristiques associées aux termes. Les auteurs utilisent également une mesure de similarité utilisant le cosinus. Dans (Koço et al., 2012), les *BoW* sont proposés comme des caractéristiques pour une approche de catégorisation (*AdaBoost*) bâtie sur plusieurs vues. Cinq vues séparées sont introduites pour respectivement l'agent, l'utilisateur, les frontières des tours de paroles dans le dialogue, la durée, ainsi que les entités nommées. Les expérimentations reportées ici reposent sur une augmentation du corpus de test, contenant des données complexes et diversifiées.

Le tableau 5.3 montre les résultats obtenus avec le système TF-IDF, M4D, ainsi que les deux systèmes s'appuyant sur les quaternions LRQ et SSQ. Il est à observer que la segmentation des dialogues permet une amélioration de la précision : le système M4D obtient de meilleurs résultats que le système s'appuyant sur le TF-IDF avec un gain d'environ 7 % en absolu, validant l'intuition initiale que la structure d'une conversation doit être prise en compte lors d'une tâche de catégorisation de documents bruités.

Les quaternions fournissent un gain absolu additionnel de 6,5 % pour la segmentation LRQ et 8,6 % pour la segmentation SSQ. Ces résultats confirment que, au contraire des représentations matricielles classiques comme le système M4D, les quaternions permettent de capturer les dépendances liant les distributions des termes composant le

document. Ces dépendances semblent clairement pertinentes lors d'une tâche d'extraction de thèmes.

La différence entre les segmentations LRQ et SSQ est limitée mais relativement inattendue. Ce résultat peut être dû au fait que la structure latente des conversations peut ne pas correspondre aussi bien qu'attendu au schéma proposé en quatre segments. La comparaison des systèmes fondés sur un seul segment (voir tableau 5.2) montre un léger avantage pour le premier segment, mais les performances individuelles sont bien éloignées de celles obtenues avec une combinaison des caractéristiques (M4D, LRQ et SSQ). Globalement, ces résultats suggèrent que les autres approches de segmentation peuvent encore améliorer les performances globales du système.

Finalement, l'utilisation de documents issus du corpus *gold* (TMAN) et de transcriptions automatiques (TRAP) (TMAN+TRAP) dans le tableau 5.3, montre des améliorations significatives, en comparaison aux KNN appliqués à une des deux sources de transcription (TRAP ou TMAN). La combinaison des deux ensembles de données (TMAN+TRAP) lors de la phase d'apprentissage, augmente la robustesse du système de reconnaissance de la parole aux erreurs de transcription.

5.5 Conclusions sur l'apport d'une représentation d'un document bruité dans un espace hyper-complexe

La représentation structurée fondée sur les quaternions a montré son apport lors d'une tâche de catégorisation. En effet, la structuration de tels documents doit être prise en compte lorsque l'on sait qu'un dialogue peut contenir plus d'un seul thème. Ces sujets peuvent être abordés à des instants différents de la conversation. La relation latente liant les distributions des termes composant le document est une donnée essentielle car elle permet de localiser les segments traitant d'un sujet ainsi que de mesurer l'importance de ces sujets dans chacun des segments.

Les résultats obtenus ont validé aussi bien l'idée de capturer les variations tout au long de la conversation, que la pertinence du codage d'un ensemble de caractéristiques dans des quaternions. Lors de la comparaison des différentes représentations matricielles de grandes dimensions, les quaternions ont montré qu'ils permettent de modéliser efficacement les dépendances entres les caractéristiques, en utilisant des mesures fondées sur des distortions minimales entre conversations. Les résultats obtenus ont montré un gain significatif. Les caractéristiques bâties sur les quaternions montrent également que les dépendances entre les caractéristiques sont fortes.

Ce travail devrait être poursuivi en :
— Évaluant la représentation du document par un vecteur de quaternions sur d'autres type de documents (article Wikipédia, documents issus du Web ...)
— En utilisant des méthodes de classifications plus élaborées et tenant compte de la structure temporelle du document.

Quatrième partie

Conclusions et Perspectives de recherche

Conclusions générales

Le développement d'Internet a conduit à une croissance très forte du nombre de documents disponibles sur la toile ainsi que de la popularité des sites de partage comme Twitter ou Youtube. Par ailleurs, l'utilisateur du Web est devenu un acteur dans l'élaboration et la diffusion de documents très variés. Cette diversité porte sur les supports (texte, audio, et/ou image), sur la forme des documents (contraintes de taille, style d'écriture, etc.), sur la nature des sujets abordés et la façon dont ils sont abordés .... Le traitement d'une telle masse d'information requiert l'utilisation de systèmes fiables, performants, rapides et robustes à ces variabilités.

Les systèmes d'analyse des contenus nécessitent, au préalable, de représenter les données dans des espaces vectoriels relativement compacts dans lesquels l'information utile est facilement accessible. Plusieurs approches, détaillées dans le chapitre 1, considèrent le document comme un ensemble de termes sans tenir compte de l'ordre des mots. Représenter les documents par des vecteurs de fréquence de mots est l'approche la plus classique dans le domaine de la recherche d'information, mais ce modèle vectoriel souffre de deux défauts majeurs :

— d'une part, il s'agit d'une représentation au niveau lexical qui peut être fortement dégradée par des altérations de la forme de surface des documents (par exemple lorsqu'il s'agit de SMS, de *tweets* ou de textes produits par un système de transcription automatique),

— d'autre part, en considérant le contenu lexical d'un document comme un sac-de-mots ou comme un vecteur de fréquence de mots, nous perdons la structure temporelle du document, structure qui porte une partie des contenus sémantiques.

Dans ce mémoire, nous nous sommes attaqués à ces deux problèmes de fond liés à la représentation de documents textuels ou parlés.

La première proposition repose sur l'idée que la projection des documents dans des espaces thématiques devrait permettre de limiter l'impact du bruit observé au niveau lexical. Les espaces thématiques, comme LSA, PLSA, ou LDA, permettent de représenter des documents dans un espace abstrait de classes (ou concepts). Le chapitre 2 présente plusieurs études montrant l'apport d'une représentation de documents bruités issus de plateformes de partage ou de systèmes de reconnaissance automatique de la parole, dans un espace de thèmes LDA. Pour mesurer la performance de telles représentations de haut niveau, celles-ci sont comparées à une représentation classique

utilisant la fréquence de mots lors de tâches de contextualisation de messages courts, d'extraction d'information depuis des vidéos, ainsi que de catégorisation de transcriptions automatiques. Lors de ces expérimentations, les résultats obtenus lorsqu'une représentation thématique est utilisée, sont bien meilleurs que la représentation du document uniquement par la fréquence des termes qui le composent, avec un gain de plus de 10 points lors de la tâche de catégorisation de conversations issues de transcriptions automatiques.

Ces expérimentations ont été conduites avec des espaces de thèmes LDA dont les hyper-paramètres ont été choisis empiriquement. Ce choix constitue une force du modèle, car il permet de "régler" le modèle à la tâche pour laquelle ces espaces sont estimés. Dans le même temps, le choix est souvent épineux, car il n'existe pas de règles claires permettant de choisir les valeurs de ces hyper-paramètres de manière efficace et universelle. Le paramètre le plus important est le nombre de classes composant le modèle. Pour pallier cette difficulté, le chapitre 3 propose de représenter un document bruité dans plusieurs espaces de thèmes. Les expérimentations sont menées sur une tâche de détection automatique d'événements dans une base de données d'images issue de la campagne MediaEval 2011 SED, ainsi que sur une tâche de catégorisation de documents bruités issus de transcriptions automatiques. Les gains observés sont encourageants, avec respectivement $1,7$ et $1,6$ points de mieux lors des tâches de détection et de catégorisation de dialogues, mais la combinaison des vues est basique car elle repose sur un simple vote entre classifieurs.

La projection d'un document bruité dans plusieurs espaces de thèmes permet d'obtenir des vues complémentaires sur le document mais elle présente aussi des inconvénients : la taille de l'espace de représentation est considérablement augmentée, les vues sont très redondantes, une nouvelle variabilité est ajoutée et cette variabilité n'est pas liée aux données mais à la diversité des vues. Enfin, cette représentation n'est pas une fin en soi et ces vues devront être combinées dans un module d'analyse des contenus.

La réduction des variabilités nuisibles est un problème qui a été très exploré dans le domaine de la vérification du locuteur, dans lequel un même locuteur peut être enregistré dans des conditions et dans des environnements différents. Ces variabilités liées à la session et à l'environnement s'ajoutent à celle de la voix du locuteur lui-même. Pour compenser ces variabilités, les chercheurs du domaine ont proposé un ensemble de méthodes issues de l'analyse factorielle jointe (JFA).

Nous avons utilisé ces méthodes pour la réduction des représentations multi-vues d'un même document dans des espaces de thèmes. Dans le chapitre 4, cette idée est développée en proposant une méthode associant une représentation multi-vues et une technique d'extraction du sous-espace pertinent par analyse factorielle. La première étape de cette étude a consisté à évaluer la dépendance des modèles LDA à ses hyper-paramètres (nombre de classes et paramètres de contrôle des distributions). Cette étude nous a amené à obtenir des vues en faisant varier ces paramètres. La diversité des vues est ensuite réduite par une méthode issue de la JFA, fondée sur les i-vecteurs.

Finalement, la méthode proposée permet de s'affranchir du choix des hyper-paramètres des modèles LDA, de produire des vues complémentaires d'un document

et de tirer le meilleur parti de cette complémentarité en réduisant la représentation par analyse factorielle et non par un simple vote entre divers classifieurs. Nos expériences, menées sur un corpus de dialogue humain-humain et sur une tâche de classification de documents textuels, ont montré l'intérêt de ce mécanisme d'expansion/compression de la représentation.

Le modèle génératif probabiliste LDA considère le document comme un sac-de-mots. Ainsi, l'ordre aussi bien que la position des termes dans le document ne sont pas pris en compte lors du processus de modélisation des espaces de thèmes LDA. Un même document peut contenir plus d'un seul sujet, mais ceux-ci sont abordés de manière plus ou moins importante. Ainsi, dans un même document, des sujets multiples peuvent être abordés, et ceux-ci se trouvent alors concentrés dans des zones conversationnelles précises et différentes. Pour cette raison, la prise en compte de la position d'un terme au sein du document est une information toute aussi importante que le nombre d'occurrences de ce terme dans ce document. Le chapitre 5 étudie l'apport d'une représentation permettant de tenir compte de la position des termes au sein du document. Pour ce faire, un hyper-complexe de dimension quatre, appelé quaternion, est utilisé pour coder les fréquences d'un mot dans quatre zones ou instants différents d'une conversation. Lors de la tâche de catégorisation de documents, les résultats obtenus avec une représentation dans l'espace des quaternions sont supérieurs de $6,3$ points comparativement à une représentation matricielle comportant les mêmes caractéristiques issues des quatre segments. Ces résultats permettent de confirmer l'intuition que la position des termes au sein de documents structurés est une information importante à prendre en compte lors d'une tâche de catégorisation.

Finalement, les travaux de recherche menés ont proposés deux idées originales pour la représentation de textes bruités. La première est basée sur une approche multi-vues combinée à une méthode de fusion utilisant l'espace des *i*-vecteurs issu de l'analyse factorielle. La seconde intègre des informations liées à la structure temporelle du document dans un nombre hyper-complexe appelé *quaternion*. La représentation la plus efficace, robuste, et indépendante des paramètres propres aux modèles LDA, est la représentation fondée sur l'espace des *i*-vecteurs. Cette représentation contient essentiellement (mais pas uniquement) une information utile et robuste permettant de catégoriser efficacement des documents fortement bruités.

Perspectives

Les travaux présentés dans ce manuscrit ont abouti à une représentation robuste des documents bruités dans un espace compact de dimension réduite. Cette représentation s'appuie sur une projection multiple d'un document dans des espaces de thèmes LDA, puis sur la fusion de ces représentations thématiques pour obtenir une représentation en *i*-vecteurs. Le principe de ce mécanisme d'expansion puis fusion pourrait être reformulé et évalué pour diverses approches nécessitant une représentation multi-vues. Un obstacle éventuel à la généralisation de cette approche peut être que les performances se révèlent dépendantes de la nature des vues qui sont produites ; en effet, l'analyse factorielle repose sur l'hypothèse que les variabilités indésirables peuvent être isolées dans un sous-espace de dimension réduite, ce qui est difficile à vérifier *a priori*.

Le chapitre 5 utilise les espaces hyper-complexes pour introduire de l'information temporelle dans les modèles. Ceci permet de localiser l'apparition des thèmes lors d'une conversation. La méthode de classification utilisée lors des expérimentations est la recherche des k plus proches voisins. L'utilisation d'une méthode plus performante, comme les réseaux de neurones de quaternions, devrait permettre de tirer un meilleur bénéfice d'une telle représentation.

Les représentations des documents traités dans ce manuscrit s'appuient sur des espaces de thèmes LDA. Lors de l'apprentissage de ces modèles, l'information liée à l'étiquette associée à la conversation n'est pas prise en compte. En effet, cette information est utilisée uniquement lorsque les modèles sont estimés, durant la phase d'apprentissage des classifieurs. Le modèle *Author-Topic Model* (AT) (Rosen-Zvi et al., 2004) permet de considérer, en plus du contenu lexical du document (*i.e.* les mots qui le composent), un ensemble d'auteurs. Ainsi, la notion d'auteur pourrait être généralisée à une caractéristique quelconque du document ou a une distribution de propriétés. Ce processus permettrait de considérer à la fois les termes composant le document et, dans le même temps, les étiquettes lors de la phase d'apprentissage des espaces de thèmes. Ce modèle permet alors d'associer à chacune des classes AT-LDA une distribution sur l'ensemble de termes, mais en plus, une distribution sur toutes les étiquettes. D'autres modèles permettent de considérer l'étiquette dans le processus d'apprentissage des espaces de thèmes comme la méthode *supervised LDA* (Blei et McAuliffe, 2007) ou la méthode *labeled LDA* (Ramage et al., 2009), qui permettent d'obtenir directement une distribution sur les étiquettes lorsque l'on infère un nouveau document provenant du corpus de validation.

D'une façon générale, nous avons considéré le problème de la représentation indépendamment des modules "clients" qui utiliseront ces représentations pour réaliser des analyses dépendantes de la tâche visée. L'intégration et l'optimisation conjointe des différents niveaux est une voie qui nous semble prometteuse. Enfin, les deux principaux modèles originaux que nous avons proposés pourraient être reconsidérés dans la perspective d'une intégration des modules d'analyse, par exemple par l'utilisation de réseaux profonds de quaternions ou par une fusion des différentes vues guidée par la tâche.

Annexes

Annexe A

Modèle à base de fréquences de mots

Cette annexe décrit le critère TF-IDF, le dérivé Okapi et présente un exemple concret d'utilisation.

A.1 Formulation du TF-IDF

Le TF-IDF est un critère évaluant le poids sémantique d'un mot dans un document. Pour un mot ou terme t_i contenu dans le document d_j du corpus C, le TF-IDF de t_i est donné par :

$$tf(t_i, d_j) = \frac{f(t_i, d_j)}{|d_j|}, \qquad idf(t_i) = log\frac{|C|}{n(t_i)}, \qquad n(t_i) = |\{d_j \in C : t_i \in d_j\}| \tag{A.1}$$

avec $f(t_i, d_j)$ la fréquence du terme t_i au sein du document d_j, $|C|$ le nombre de documents contenus dans le corpus C, et $n(t_i)$ le nombre de documents où le terme t_i apparaît.

Ainsi, un mot apparaissant dans tous les documents du corpus aura un *idf* nul. Au contraire, un mot rarement présent dans les documents aura, lui, un *idf* très élevé. L'*idf* est généralement associé au logarithme de l'inverse de la fréquence du mot au sein du document, suivie d'une phase de normalisation comme pour le calcul du *tf*. Après le calcul du TF-IDF de chacun des mots contenus dans chaque document composant le corpus, une matrice A terme-par-document est obtenue. Les colonnes de celle-ci représentent le TF-IDF de chaque mot du document. Ainsi, l'espace de représentation du document est de taille fixe et connue, et cette taille est le nombre de mots contenus dans le vocabulaire V.

Une version plus récente du TF-IDF, appelée Okapi ou BM25 (Robertson et al., 2000), a montré son efficacité pour la première fois lors de la campagne d'évaluation

TREC 2010 (Text REtrieval Conference [1]). Celle-ci est déclinée sous plusieurs formes dont la plus communément utilisée est la suivante :

$$tf(t_i, d_j) = \frac{f(t_i, d_j).(k+1)}{f(t_i, d_j) + k.(1 - b + d.\frac{|d_j|}{\overline{dl}})}, \qquad idf(t_i) = log\frac{|C| - n(t_i) + 0,5}{n(t_i) + 0,5} \tag{A.2}$$

avec $\overline{dl}$ la taille moyenne d'un document contenu dans le corpus C, k et b étant des paramètres habituellement fixés à $k \in [1,2;2,0]$ et $b = 0,75$, ceci en l'absence d'une méthode efficace d'optimisation permettant d'attribuer des valeurs optimales à ces paramètres. Une étude approfondie comparant les différentes déclinaisons du TF-IDF peut être trouvée dans (Savoy et Dolamic, 2008).

A.2 Matrice de représentation

Les deux indicateurs du pouvoir discriminant d'un terme (TF-IDF) sont généralement associés pour constituer chacun des composants $a_{i,j}$ de la matrice A :

$$a_{i,j} = tf(t_i, d_j) \times idf(t_i) \tag{A.3}$$

Depuis cette première écriture du coefficient $a_(i,j)$, plusieurs déclinaisons ont vu le jour, comme la pondération normalisée par le cosinus calculée selon la formule suivante (Salton et Buckley, 1988) :

$$a_{i,j} = \frac{tf(t_i, d_j) \times idf(t_i)}{\sqrt{\sum_{k=1}^{|V|} tf(t_k, d_j) \times idf(t_k)}} \tag{A.4}$$

D'autres caractéristiques d'un terme sont également utilisées comme le détaille la section A.3.

A.3 Autres caractéristiques d'un terme dans un document

Les deux descripteurs d'un terme au sein d'un document permettent de mesurer son nombre d'apparitions dans le corpus (*tf*), et le nombre de documents le contenant (*idf*). Il est fréquent que l'on adjoigne d'autres caractéristiques du terme au sein du corpus de documents, selon la tâche dont cette représentation est dévolue, et selon la taille et la variété des documents composant ce corpus. Un critère efficace adjoint habituellement dans le calcul de la caractéristique représentant le terme dans les tâches d'extraction de mots-clés par exemple, est le critère de pureté de Gini détaillé dans la section A.3.1. La position relative d'un terme au sein d'un document est également souvent utilisée dans des tâches où l'ordre des termes ou des phrases a une importance, comme le résumé automatique. Ce dernier descripteur est présenté dans la section A.3.2.

1. http ://trec.nist.gov/

A.3.1 Critère de pureté de Gini

Le pouvoir discriminant d'un terme est donc donné par la valeur de TF-IDF qui lui est associée. Un critère permettant de déterminer le pouvoir discriminant inter-classes, est le critère de pureté de Gini (Demiriz et al., 1999). Ce critère favorise les termes apparaissant souvent dans une classe, tout en étant rares ou peu apparents dans toutes les autres classes. Il est défini comme suit :

$$gini_t(w) = 1 - \sqrt{\sum_{i=1}^{|\mathbf{T}|} p_i^2} \, ,$$

où p_i est la probabilité que le terme w soit généré par la i^{me} classe, et $t \in \mathbf{T}$ est l'ensemble des classes appartenant au corpus.

A.3.2 Position relative d'un terme

Ces trois indicateurs ont été rejoints par un troisième (rp) décrivant la position du mot dans le document (*tf-idf-rp*). Celui-ci a pour vocation de donner un poids plus important aux mots apparaissant pour la première fois en début de document. Il est défini ainsi dans (Xi et al., 2013) :

$$rp(t_i, d_j) = \frac{1 - pos(t_i, d_j)}{|d_j|} \tag{A.5}$$

où $pos(t_i, d_j)$ est le nombre d'occurrences de mots apparaissant avant le terme t_i. Dans (Liu et al., 2008; Morchid et Linarès, 2013b) la position relative est définie selon l'équation :

$$rp(t_i, d_j) = \frac{1}{pos(t_i, d_j)} \tag{A.6}$$

A.4 Exemple : Mots discriminants et TF-IDF

Cette section présente un exemple d'utilisation du TF-IDF dans une tâche d'extraction d'un ensemble de mots-clés permettant de discriminer, au mieux, un ensemble de documents. Celui-ci est composé de cinq articles traitant de trois sujets bien distincts : le football (Olympique de Marseille, Olympique Lyonnais et Paris-Saint-Germain), la politique (François Hollande), et un fait majeur historique (seconde guerre mondiale). Le tableau A.1 détaille les caractéristiques de ces documents composés d'un vocabulaire de 247 mots. Ces documents ont été pré-traités en supprimant les mots usuels (de, du, les, ...).

La figure A.1 montre un ensemble (7) de mots discriminants rangés dans l'ordre décroissant du haut vers le bas, plus un mot (français) très peu discriminant (*tf-idf=0.007*). La première remarque que l'on peut faire, concerne la corrélation entre le score obtenu

Sujet du document d_j	$\lvert d_j \rvert$	# mots uniques dans d_j	% mots uniques dans d_j
François Hollande	69	56	81,16
Olympique Lyonnais	50	41	82,00
Olympique de Marseille	57	43	75,44
Paris-Saint-Germain	70	54	77,14
Seconde Guerre Mondiale	119	85	71,43

TABLE *A.1* – *Description du corpus.*

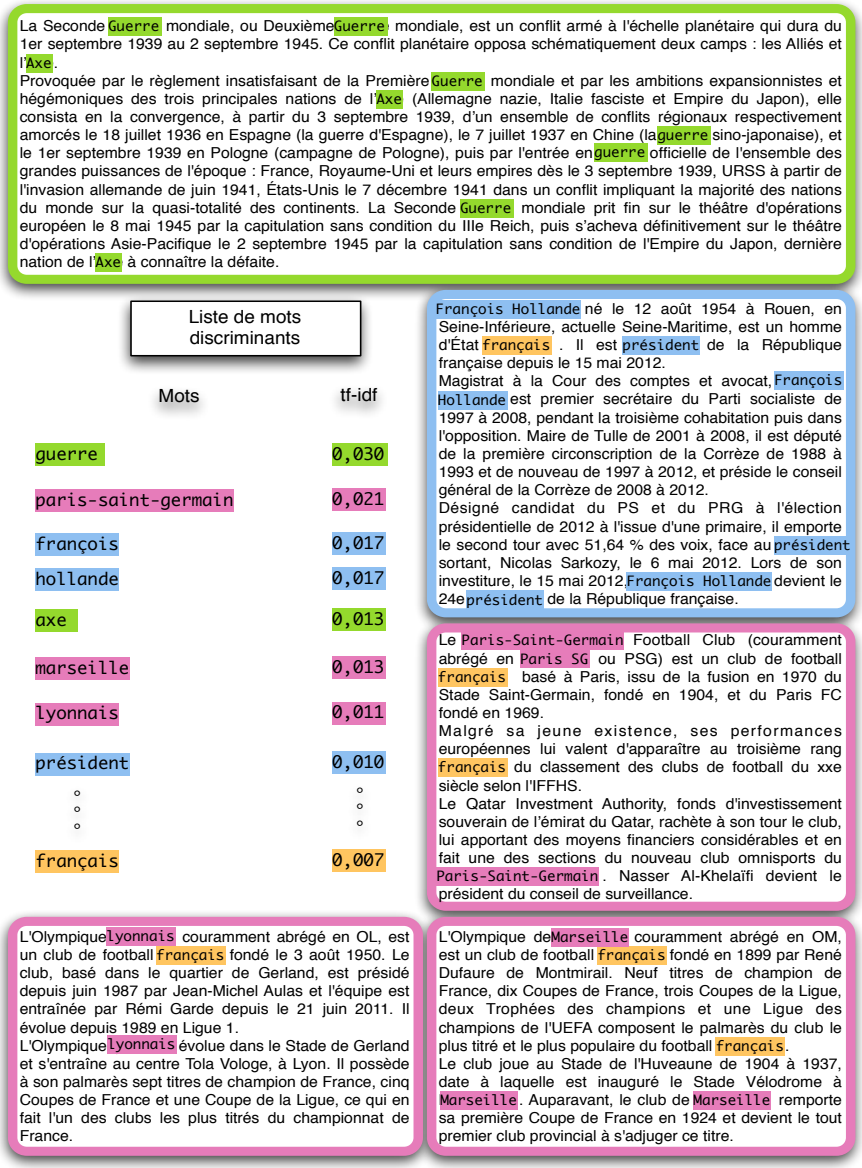


FIGURE *A.1* – *Illustration du pouvoir discriminant du tf-idf.*

par les mots, en termes de TF-IDF, et leur propension à appartenir à très peu de documents. Les mots *guerre* et *axe* apparaissent uniquement dans le document traitant du sujet *Seconde guerre mondiale* (encadré vert). Au contraire de ces mots au pouvoir discriminant important, nous trouvons un mot plus générique et ne permettant pas de distinguer un document d'un autre, le mot *français*. Celui-ci apparaît dans la majorité des documents ($\frac{4}{6}$ surligné en brun sur la figure A.1).

Annexe B

Analyse sémantique latente (LSA)

Dans le chapitre 2, l'analyse sémantique latente est introduite comme une méthode de réduction de l'espace de représentation terme-document. Elle permet également de visualiser un document de la collection comme une combinaison de thèmes latents. Les sections suivantes explicitent, dans un premier temps, le procédé de réduction (SVD) puis son utilisation dans le cadre de la recherche d'information et de matrice terme-document (LSA/glelsi) (Bellegarda, 1997, 2000).

B.1 Décomposition en valeurs singulières (SVD)

Cette section décrit, en détail, la réduction de l'espace de représentation d'une matrice terme-document par la décomposition en valeurs singulières est communément utilisée comme une solution au problème de moindre carré non-contraint, à l'estimation de rang de matrice, ou encore à l'analyse de corrélation canonique (Berry, 1992). Cette technique est proche de la décomposition en valeurs propres et de l'analyse factorielle (*factor analysis*) (Cullum et Willoughby, 1985).

Avant tout, nous présentons les notations nécessaires comme suit :

— A : matrice de dimension $m \times n$ avec $m \geq n$ et $\text{rang}(A) = r$.
— I_n : matrice identité d'ordre n.
— $R(A)$: rang de la matrice A, est l'ensemble des vecteurs b pour lequel il existe un vecteur x tel que : $Ax = b$ et $\dim(R(A)) = \text{rang}(A) = r$.
— $N(A)$: espace nul de la matrice A, est l'ensemble des vecteurs x tel que : $Ax = 0$.
— U : matrice de vecteurs singuliers à gauche de dimension $m \times m$ telle que $U = [u_1 u_2 \dots u_m]$.
— V : matrice de vecteurs singuliers à droite de dimension $n \times n$ telle que $V = [v_1 v_2 \dots v_n]$.
— Σ : matrice de valeurs singulières α_i contenues dans sa diagonale et de dimension $m \times n$ telle que $\Sigma = \text{diag}(\alpha_1, \dots, \alpha_r)$ avec $\alpha_i > 0$ pour $1 \leq i \leq r$, $\alpha_j = 0$ pour $j \geq r + 1$.

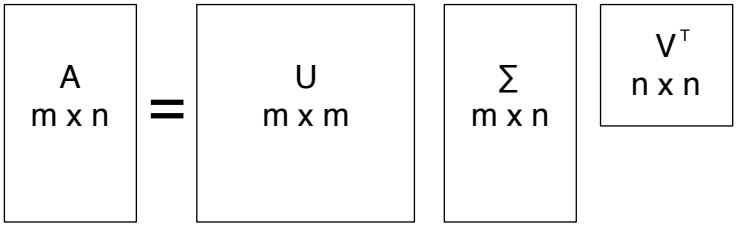


FIGURE B.1 – *Décomposition SVD d'une matrice A de dimension $m \times n$.*

La décomposition en valeurs singulières d'une matrice A, présentée dans la figure B.1, est définie comme suit :

$$A = U\Sigma V^T \tag{B.1}$$

avec $U^TU = V^TV$. Les r premières colonnes de la matrice U et V représentent les vecteurs propres associés aux r valeurs propres de AA^T et A^TA respectivement. Les éléments situés sur la diagonale de la matrice Σ sont les racines carrées positives des valeurs propres de AA^T (Golub et Van Loan, 1989). Pour illustrer la capacité de la SVD à extraire de l'importance sur la structure d'un ensemble de données sous la forme matricielle, comme la matrice A, deux théorèmes sont présentés.

THEOREME 2.1. *Soit la SVD de la matrice A définie dans l'équation B.1 et*

$$\alpha_1 \geq \alpha_2 \ldots \alpha_r \geq \alpha_{r+1} = \cdots = \alpha_n = 0, \tag{B.2}$$

Puis,

- la propriété du rang : $\text{rang}(A) = r$, $N(A) \equiv vect\{v_{r+1}, \ldots, v_n\}$ et $R(A) \equiv vect\{u_1, \ldots, u_r\}$
- la décomposition dyadique de la matrice A : $A = \sum_{i=1}^{r} u_i.\alpha_i.v_i^T$.

La preuve de ce théorème peut être trouvée dans (Golub et Van Loan, 1989).

THEOREME 2.2. *Soit la SVD de la matrice A donnée par (B.1) avec $r = rang(A) \leq p = min(m, n)$, où est défini :*

$$A = \sum_{i=1}^{r} u_i.\alpha_i.v_i^T \tag{B.3}$$

alors

$$\min_{\text{rang}(b)} = ||A - B||_F^2 = ||A - A_k||_F^2 = \alpha_{k+1}^2 + \cdots + \alpha_p^2 \tag{B.4}$$

La preuve de ce théorème peut être trouvée dans (Wilkinson et al., 1971).

Dans d'autres termes, A_k, qui est construite à partir des k plus grandes valeurs singulières contenues dans la diagonale de la matrice Σ, est la plus proche matrice de rang k de la matrice originale A. La matrice A_k représente alors la meilleure approximation de la matrice A pour n'importe quelle norme unitaire invariante (Mirsky, 1960). Ainsi :

$$\min_{\text{rang}(b)} = ||A - B||_2 = ||A - A_k||_2 = \alpha_{k+1}. \tag{B.5}$$

Pour plus de détails sur le modèle mathématique, voir (Furnas et al., 1988).

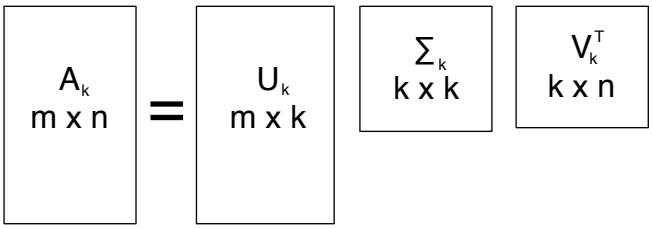


FIGURE B.2 – *Réduction par la méthode SVD d'une matrice A de dimension $m \times n$ vers une matrice A^k de dimension $m \times k$.*

La matrice A_k est donc une approximation de la matrice A dans un espace réduit ayant pour dimension les k plus grandes valeurs singulières contenues dans la matrice Σ de dimension $k \times k$, comme indiqué dans la figure B.2.

B.2 Application de LSA pour le traitement de l'image

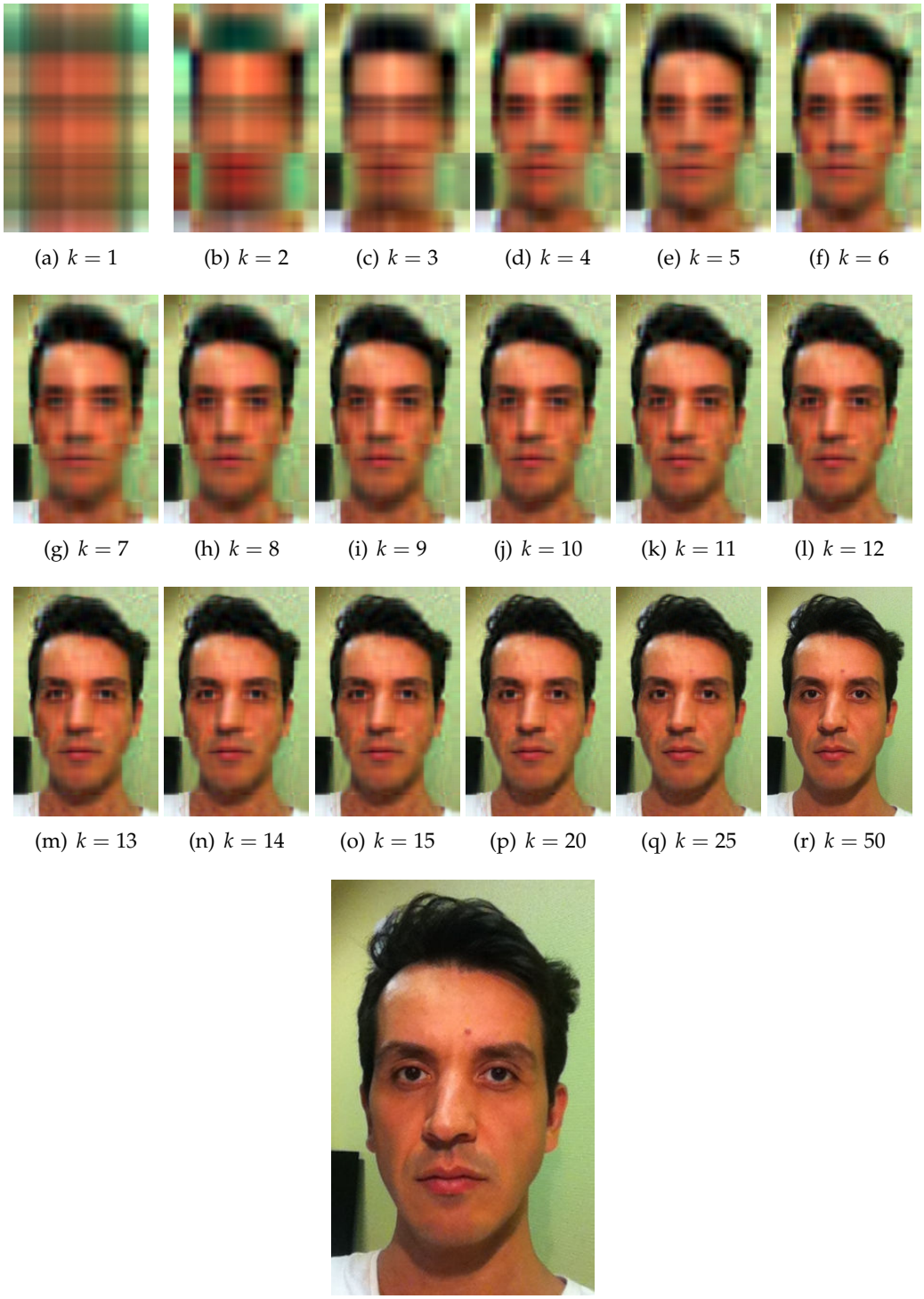
La section précédente a montré que la matrice A_k ($k \leq \text{rang}(A)$) est la matrice la plus proche de A. Plus la valeur de k est petite (proche de 0), plus la matrice A_k diffère de la matrice "originale" A. Pour illustrer le théorème 2.2, une image est projetée dans un espace RGB (*Right Green Blue*) pour obtenir une matrice de pixels (voir figure B.3). Celle-ci est ensuite décomposée en valeurs singulières. Une nouvelle matrice A_k est, par la suite, définie en ne conservant uniquement les k plus grandes valeurs.

Dans le cas d'une image de dimension 437×276 présentée dans la figure B.3, on observe qu'une approximation de la matrice originale A au rang $k = 50$ permet d'obtenir une image proche de l'image originale ($k = 276$) représentée par la matrice A. De plus, le format employé (JPEG) n'est pas un format vectoriel et comporte des perte quand la qualité de la matrice de restitution diffère de celle de départ.

B.3 Indexation sémantique latente

La technique d'indexation sémantique latente (Indexation Sémantique Latente ou *Latent Semantic Indexation* (LSI)) (Bellegarda, 1997, 2000) utilise la décomposition SVD, vue dans la section précédente, pour construire les relations associatives entre les mots d'une collection de documents. L'indexation sémantique latente est réalisée à partir d'une grande matrice A. Celle-ci est décomposée en k facteurs orthogonaux depuis laquelle, la matrice originale A peut être approximée par combinaison linéaire.

Au lieu de représenter les documents au sein d'un corpus comme une combinaison de mots indépendants, la technique LSI les présente comme des facteurs projetés dans l'espace des vecteurs orthogonaux dans l'espace de décomposition de la SVD. Cette



(a) $k = 1$ (b) $k = 2$ (c) $k = 3$ (d) $k = 4$ (e) $k = 5$ (f) $k = 6$

(g) $k = 7$ (h) $k = 8$ (i) $k = 9$ (j) $k = 10$ (k) $k = 11$ (l) $k = 12$

(m) $k = 13$ (n) $k = 14$ (o) $k = 15$ (p) $k = 20$ (q) $k = 25$ (r) $k = 50$

(s) Image originale

FIGURE B.3 – *Décomposition SVD appliquée à une image pour* $1 \leq k \leq 50$ *(*$A = [437 \times 276]$*).*

espace étant de taille réduite comparativement à l'espace d'origine du vocabulaire, les mots perdent ainsi leur caractère indépendant. Si, par exemple, deux mots sont utilisés dans des contextes identiques (documents), ils auront un vecteur de représentation identique dans l'espace réduit LSI.

Cette méthode permet alors de dépasser certaines faiblesses du modèle simple s'appuyant sur la fréquence de mots (*tf-idf*) en annihilant partiellement l'aspect indépendant des termes, et ainsi mettre en évidence la relation latente terme-terme, document-terme, et document-document, qui est régie par la distribution des mots au sein des documents.

Dans l'annexe A, la matrice A était définie à l'aide de poids indiquant le caractère discriminant d'un terme au sein d'un document appartenant à un corpus (ou collection de documents). Ainsi, la matrice A est définie comme suit :

$$A = [a_{i,j}] \tag{B.6}$$

avec $a(i, j)$ défini comme dans (A.3) pour une version du *tf-idf* non-normalisée, ou (A.4) pour une normalisation à l'aide d'un cosinus.

B.4 Espace sémantique LSA

La section précédente décrit la méthode de réduction de l'espace de représentation d'un corpus de documents dans un espace de représentation thématique à l'aide d'une décomposition en valeurs singulières (SVD). Une illustration de cette technique est présentée dans cette section en utilisant le corpus de documents décrit à la section A.4.

Ainsi, la figure B.4 présente une réduction de l'espace terme-document dont la dimension correspond à la taille du vocabulaire multipliée par le nombre de documents (247×5). L'espace réduit obtenu est de dimension 247×3 ($k = 3$) avec, pour chacun des 3 *topic* (thème), les 4 mots les plus représentatifs du thème z_i, c'est-à-dire ceux dont les probabilités pour chacun des termes m du vocabulaire de taille 247 ($p(m|z_i)$) sont les plus élevées.

La première observation, est que LSA réduit bien l'espace de représentation des documents et du vocabulaire. Ainsi l'espace de représentation est l'espace thématique réduit dans lequel les documents, comme les mots composant le vocabulaire, trouvent une représentation homogène et robuste. Dans le nouvel espace thématique, un document aura alors une représentation sur les r dimensions de l'espace thématique, avec r le nombre de valeurs singulières conservées dans la diagonale de la matrice de *scale* (échelle) Σ (voir équation B.1).

Nous constatons également que la méthode LSA "rassemble" bien les termes traitant d'un même sujet dans un seul et même thème. Ainsi, les termes décrivant la *seconde guerre mondiale* apparaissent dans un thème séparé (thème 1) de celui décrivant *François Hollande* (thème 2) ou traitant du *football* (thème 3). Ce dernier thème ne contient que

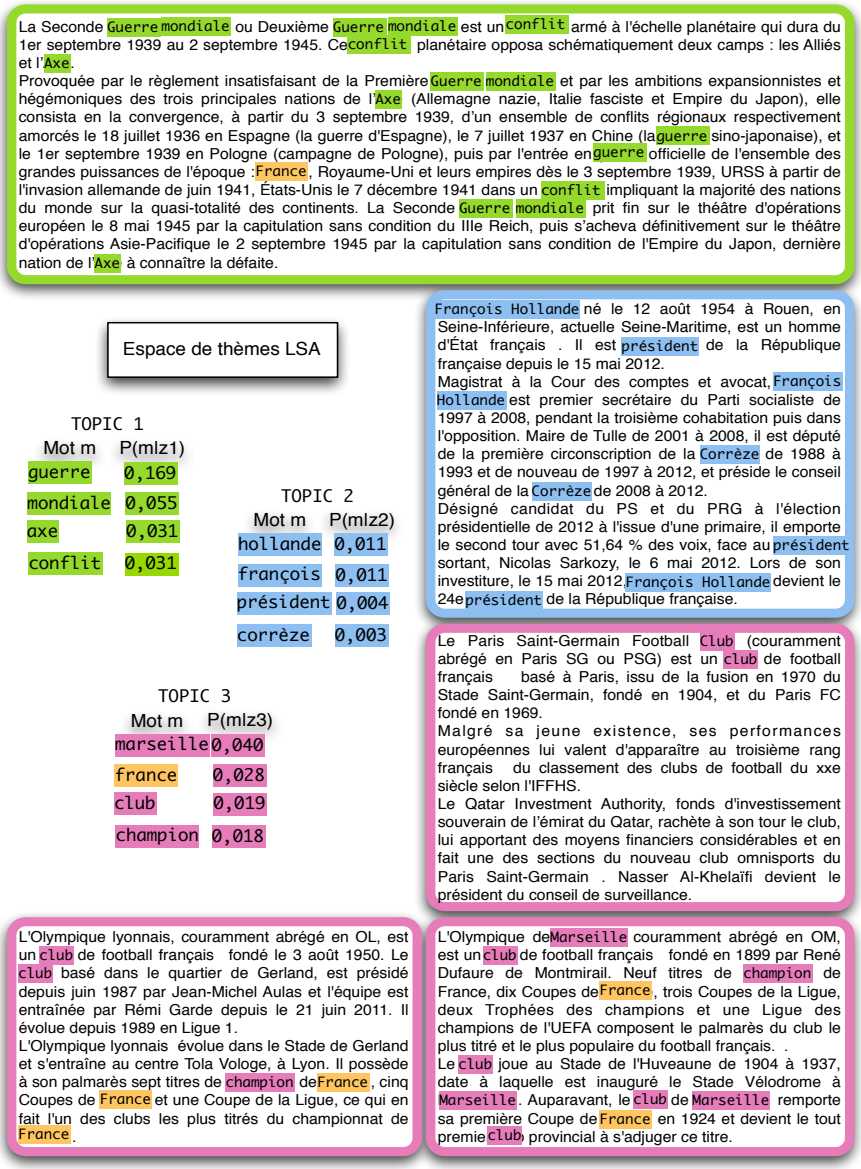


FIGURE *B.4* – *Exemple d'un espace sémantique obtenu avec la méthode LSA.*

peu de termes (le terme *club* uniquement) issus de l'article concernant le Paris-Saint-Germain. Ceci est dû au contenu même de cet article. Celui-ci traite plus de l'aspect économique (achat du club par un fonds d'investissement Qatari) que sur le football même. Ainsi, il dispose de moins de mots en communs avec les deux autres articles traitant du sujet *football* que les articles concernant l'Olympique Lyonnais et l'Olympique de Marseille. Ces deux derniers partagent un champ lexical propre au domaine footballistique qui permet aux termes les composant de fournir l'essentiel des termes du thème 3.

Un thème peut, cependant, contenir un terme plus générique comme le terme *France* (colorié en brun sur la figure B.4) apparaissant dans l'article concernant la *seconde guerre mondiale* également.

Annexe C

Analyse sémantique latente probabiliste (PLSA)

Cet annexe présente l'analyse sémantique latente probabiliste (PLSA) proposée par (Hofmann, 1999b) et introduite comme une version probabiliste de LSA dans le chapitre 2. Cette autre méthode est très utilisée également dans les tâches de recherche d'information.

C.1 Analyse sémantique latente probabiliste

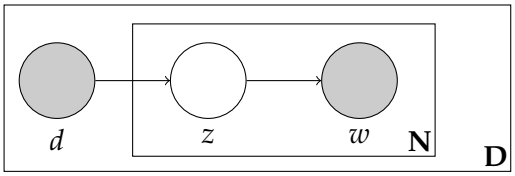


FIGURE C.1 – *Modèle PLSA.*

Le modèle PLSA fait l'hypothèse qu'un document d et un mot w sont indépendants conditionnellement sachant un thème z de l'ensemble des thèmes Z non observés comme l'indique le modèle graphique de la figure C.1. En effet, PLSA est un modèle de variables latentes permettant d'associer une classe non observée $z \in Z = \{z_1, \ldots, z_K\}$ à des données co-occurrentes comme la distribution des mots $w \in W = \{w_1, \ldots, w_N\}$ au sein d'un document $d \in D = \{d_1, \ldots, d_M\}$. Il est ainsi défini comme suit dans le sens d'un modèle génératif :

- Choisir un document d ayant une probabilité $P(d)$.
- Sélectionner une classe s de probabilité $P(z|d)$.
- Générer un mot w avec la probabilité $P(w|z)$.

Ce procédé génératif conduit à un modèle probabiliste tel que :

$$P(d, w) = P(d)P(w|d) \tag{C.1}$$

avec

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \tag{C.2}$$

l'équation suivante s'exprime alors en remplaçant $P(w|d)$ de l'équation C.1 par sa définition de l'équation C.2, comme suit :

$$P(d,w) = P(d) \sum_{z \in Z} P(w|z)P(z|d) \tag{C.3}$$

Ce modèle de mixtures (McLachlan et Basford, 1988) s'appuie sur deux hypothèses fortes :

- Le duo de variables observées (d, w) est généré de manière indépendante car le document est considéré comme un "sac-de-mots" (Salton, 1989b).
- Chacun des mots w est généré indépendamment du document sachant la variable latente (non observée) z.

Sachant que le nombre de variables latentes est beaucoup plus petit que la taille de variables observées au sein des documents formant le corpus de textes ($K << M$), les variables latentes z représentent un sous-espace de taille réduite où la probabilité de la variable w peut s'exprimer en fonction de d (P(w | d) dans l'équation C.2).

Ce qui différencie ce modèle génératif probabiliste des modèles de classification de documents ou des modèles non-supervisés comme les modèles bayésiens naïfs, est la distribution des mots w dans un document donné d $P(w|d)$ qui est obtenue par une combinaison convexe des probabilités $P(w|z)$ (Hofmann et al., 1999). Ainsi, les documents sont caractérisés par une distribution sur les variables latentes z par $P(z|d)$. Cette représentation du document comme une distribution et non comme une appartenance à une certaine classe, offre à ce modèle des modes de représentation plus fins et plus maniables dans des tâches de classification mais aussi de recherche de similarité entre documents ou ensemble de documents.

C.2 Paramètres du modèle

Ce modèle est donc composé de trois probabilités distinctes mais reliées par une variable latente : la distribution des mots w au sein des documents d. Les variables $P(d)$, $P(z|d)$, et $P(w|z)$, sont estimées en maximisant la vraisemblance L :

$$L = \sum_{d \in D} \sum_{w \in W} n(d,w) \log P(d,w) \tag{C.4}$$

où $n(d,w)$ est le nombre de fois où le terme w apparaît dans le document d. $P(d,w)$ peut être redéfini en modifiant la probabilité $P(z|d)$ en utilisant la loi de Bayes pour obtenir une version symétrique de la vraisemblance :

$$P(z|d) = \frac{P(d|z)P(z)}{P(d)} \,. \tag{C.5}$$

Ainsi, l'équation C.2 peut se réécrire en remplaçant $P(z|d)$ comme dans l'équation C.5 :

$$P(w|d) = \sum_{z \in Z} P(w|z) \frac{P(d|z)P(z)}{P(d)} \tag{C.6}$$

Cette écriture est identique à celle exprimée dans l'équation C.3, mais avec une paramétrisation différente. L'équation générale de PLSA (équation C.3) devient alors :

$$\begin{aligned} P(d,w) &= P(d)P(w|d) \\ &= P(d) \sum_{z \in Z} P(w|z) \frac{P(d|z)P(z)}{P(d)} \\ &= \sum_{z \in Z} P(w|z)P(d|z)P(z) \,. \end{aligned} \tag{C.7}$$

La vraisemblance L exprimée dans l'équation C.4 à maximiser, se calcule alors comme suit :

$$L = \sum_{d \in D} \sum_{w \in W} n(d,w) \log \left\{ \sum_{z \in Z} P(w|z)P(d|z)P(z) \right\} \tag{C.8}$$

C.3 Estimation avec l'algorithme EM

La vraisemblance L doit être maximisée pour obtenir une bonne estimation des paramètres ($P(w|z)$, $P(d|z)$ et $P(z)$) du modèle génératif probabiliste PLSA. La méthode standard est l'algorithme *Expectation-Maximisation* (EM) (Dempster et al., 1977). Cette méthode se décompose en une première phase d'estimation (E), où les probabilités sont calculées *a posteriori* pour une certaine variable latente z en utilisant les valeurs des paramètres actuels. Puis succède une phase de maximisation (M) dans laquelle les paramètres sont mis-à-jour sachant les probabilités *a posteriori* déterminées dans la phase (E) précédente.

Ainsi, la phase d'estimation (E) permet de calculer la probabilité qu'un terme w dans un document d soit déterminé par une variable latente z :

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in Z} P(z')P(d|z')P(w|z')} \,. \tag{C.9}$$

Puis, la probabilité $P(z|d,w)$ est utilisée pour estimer les nouveaux paramètres du modèle PLSA dans la phase de maximisation (M) :

$$P(w|z) = \frac{\sum_{d\in D} n(d,w)P(z|d,w)}{\sum_{d\in D, w'\in W} n(d,w')P(z|d,w')'} \tag{C.10a}$$

$$P(d|z) = \frac{\sum_{w\in W} n(d,w)P(z|d,w)}{\sum_{d'\in D, w\in W} n(d',w)P(z|d',w)'} \tag{C.10b}$$

$$R \equiv \sum_{d\in D, w\in W} n(d,w) \tag{C.10c}$$

$$P(z) = \frac{1}{R} n(d,w)P(z|d,w) \tag{C.10d}$$

Ces étapes d'estimation (E) (équation C.9) et de maximisation (M) (équation (C.10a)-(C.10d)) se succèdent jusqu'à ce que la vraisemblance L (équation C.8) converge vers un maximum local.

D'autres méthodes d'estimation des paramètres du modèle PLSA permettent une meilleure généralisation du modèle appris sur un corpus d'entraînement, sur des documents n'apparaissant pas dans ce corpus.

C.4 Éstimation avec l'algorithme TEM

Un modèle dont les paramètres sont estimés par l'algorithme EM et ayant une perplexité faible, ne garantit pas une bonne généralisation à des données non rencontrées dans la phase d'estimation. Ainsi, la variante de EM proposée par (Hofmann, 1999b), appelée EM tempéré (*Tempered EM* TEM), est fondée sur une régularisation de l'entropie. Cette méthode est proche de la méthode *deterministic annealing* (Rose et al., 1990). Ainsi, les auteurs dans (Hofmann, 1999b) introduisent un nouveau paramètre β dans l'équation C.9 permettant de calculer $P(z|d,w)$, appelé *inverse computational temperature*. Ainsi l'équation C.9 devient :

$$P_\beta(z|d,w) = \frac{P(z)|P(d|z)P(w|z)|^\beta}{\sum_{z'\in Z} P(z')|P(d|z')P(w|z')|^\beta} \,. \tag{C.11}$$

Lorsque $\beta = 1$, l'algorithme TEM est semblable à l'algorithme EM alors que dans le cas où $\beta < 1$, la partie contenue dans la norme $|.|^\beta$ est réduite. De plus amples informations concernant cette variante de EM peuvent être trouvées dans (Hofmann, 1999b).

Annexe D

Allocation latente de Dirichlet (LDA)

L'allocation latente de Dirichlet (LDA) est un modèle génératif probabiliste exploitant les relations parmi les mots et les concepts latents contenus dans un corpus de documents textuels (Blei et al., 2007). Le terme *concept* signifie que les distributions cachées lient les mots composant le vocabulaire et leurs occurrences au sein des documents. La section suivante décrit, plus en détails, les fondements de l'approche LDA.

D.1 Distribution de Dirichlet

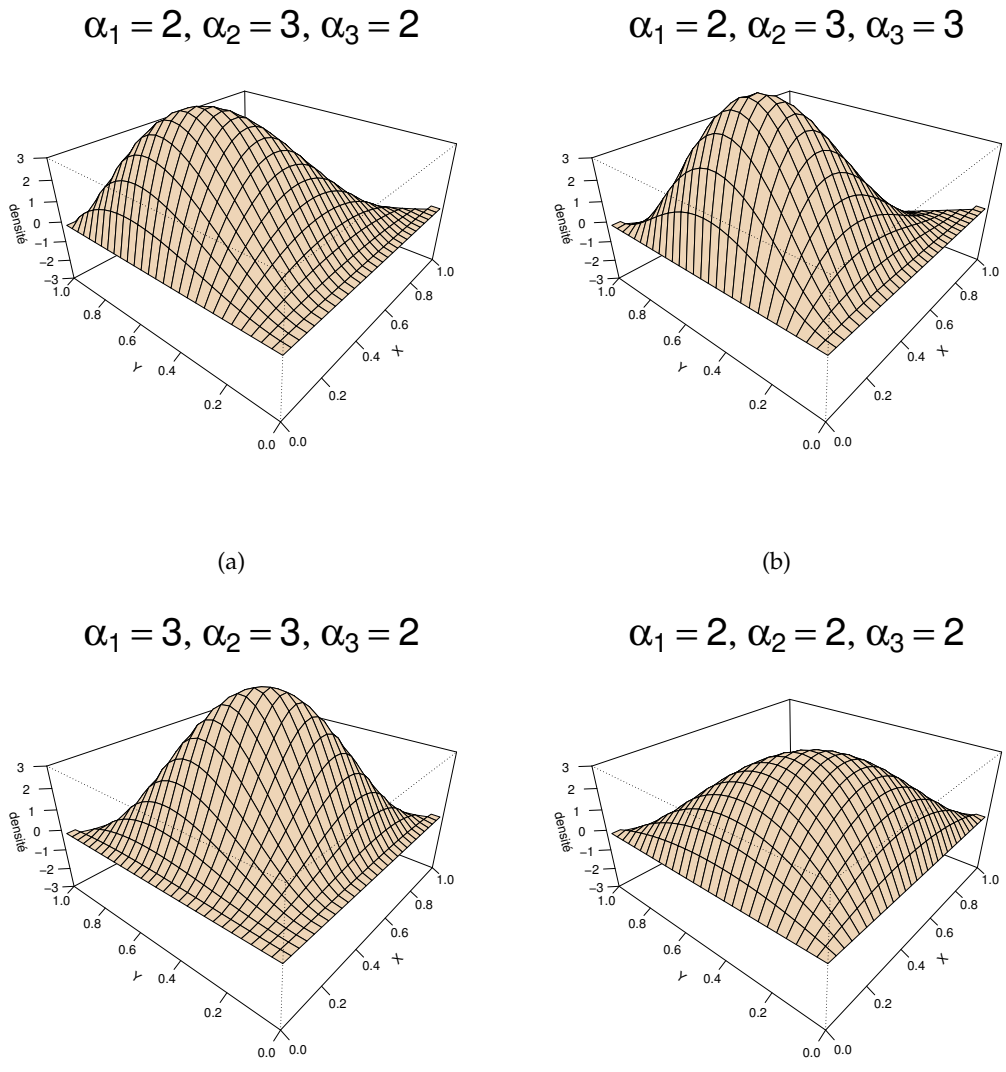
LDA est un modèle génératif probabiliste dont l'idée principale est que les caractéristiques des documents et des thèmes les composant suivent une loi de Dirichlet. Cette distribution est une généralisation multivariée de la distribution de *beta*, utilisée également dans la génération de modèles statistiques bayésiens pour la modélisation de la croyance (Neapolitan et al., 2003; Balakrishnan et Nevzorov, 2004). La distribution de Dirichlet possède une densité de probabilité (Blei et Lafferty, 2009) définie comme suit :

$$P(x|\alpha_1,\ldots,\alpha_K) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}x_i^{\alpha_i-1} \tag{D.1}$$

où α est un vecteur de dimension K positif, et Γ est une généralisation de la fonction factorielle aux valeurs réelles et appelée fonction Gamma (Neapolitan et al., 2003) :

$$\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}dt. \tag{D.2}$$

La densité de la distribution de Dirichlet varie donc en fonction de la valeur du vecteur α. Ainsi, si cette valeur est grande, un pic de densité (voir figure D.1) est observé (Blei et Lafferty, 2009). Au contraire, si la valeur de α est en deçà de 1, une densité



(a) (b)

(c) (d)

FIGURE *D.1* – *Densité de la distribution de Dirichlet pour $K = 3$ avec $\alpha > 1$.*

élevée est observée dans les parties extrêmes du n-simplex (figure D.2). Les figures D.1 et D.2 montrent différentes densités de la distribution de Dirichlet selon des valeurs du n-upplet $(\alpha_1, \alpha_2, \alpha_3)$ dans le 2-D simplex (X, Y). La figure D.2-(a) montre que la densité de la distribution de Dirichlet est uniforme pour $\alpha_1 = \alpha_2 = \alpha_3 = 1.0$, alors que la figure D.2-(b-d) indique que l'essentiel de la densité est projeté aux extrémités du simplex.

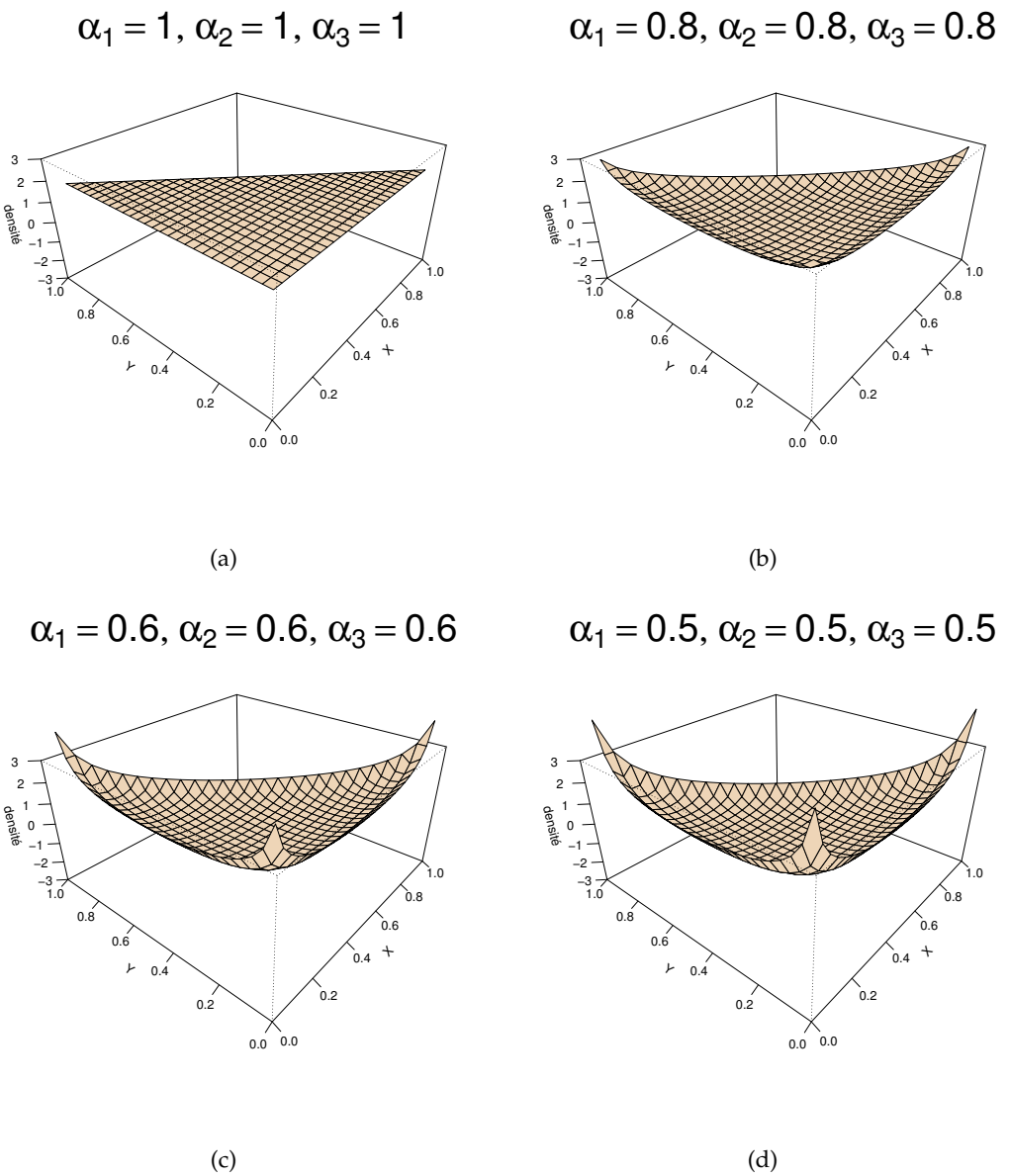


FIGURE *D.2* – *Densité de la distribution de Dirichlet pour $K = 3$ avec $\alpha \leq 1$.*

La figure montre l'influence de la valeur de α dans le calcul de la densité de la distribution de Dirichlet. Ce paramètre modifie la distribution elle-même évidement, regroupant l'essentiel de la distribution autour d'un pic, comme le montrent les figures D.1

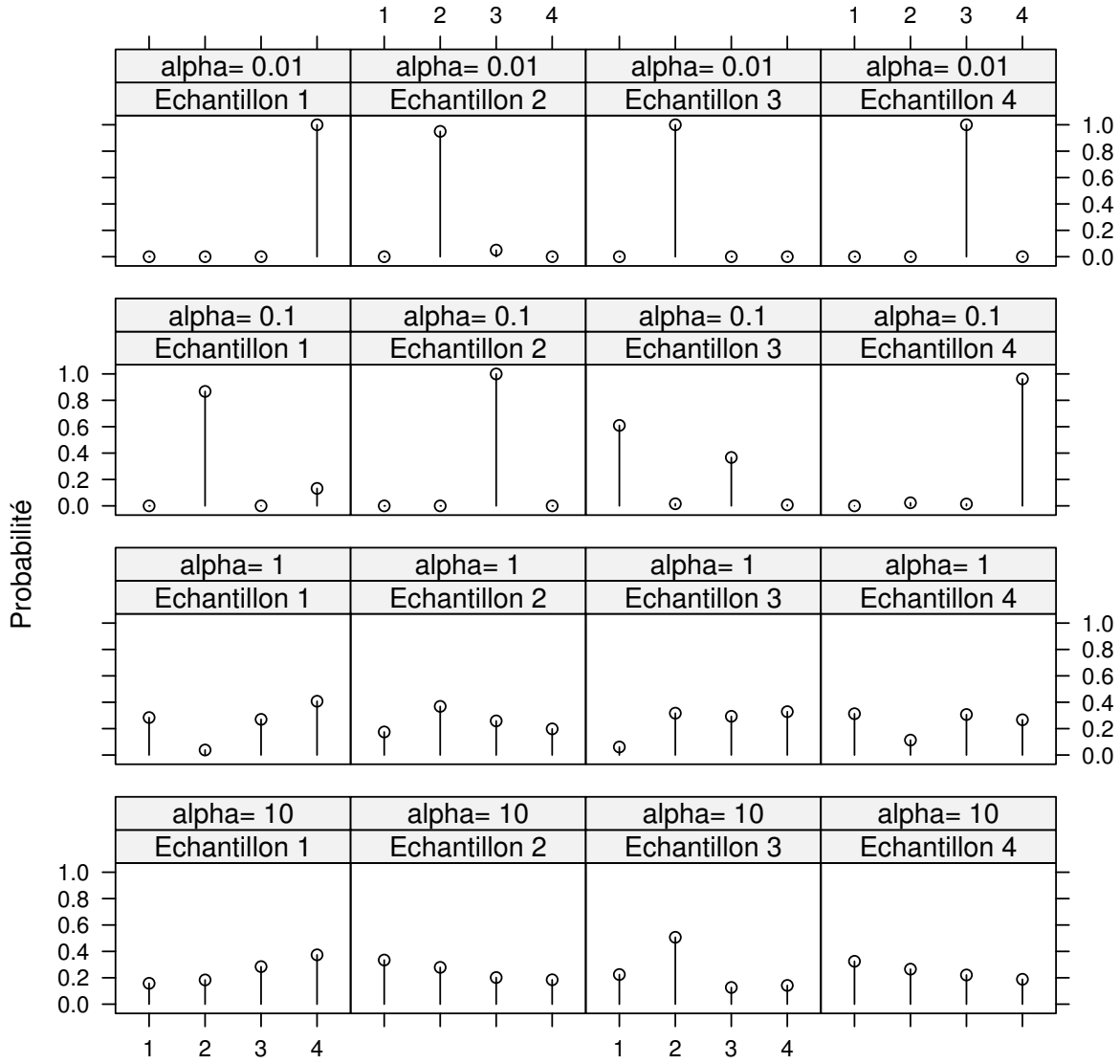


FIGURE *D.3* – *Échantillon de la distribution symétrique de Dirichlet pour $K = 4$.*

et D.3, ou, au contraire, donne un aspect de plus en plus uniforme alors que α tend vers 1.0, comme le montre la figure D.3. Les distributions de Dirichlet sont symétriques, c'est-à-dire que le vecteur α contient la même valeur pour chacune de ses composantes.

D.2 Processus génératif

Dans la description du modèle LDA, la notation contenue dans le tableau D.1 est utilisée.

D	le corpus de document
V	le vocabulaire
v	indice d'un mot du vocabulaire V
N	nombre de mots contenus dans le vocabulaire V
K	nombre de thèmes ou concepts composant le modèle
k	indice d'un thème parmi les K thèmes
d	indice sur un document
N_d	nombre de mots contenus dans le document d
i	indice d'un mot au sein d'un document
β	vecteur positif de dimension V
α	vecteur positif de dimension K
$\mathcal{D}ir(\beta)$	distribution de Dirichlet de dimension V
$\mathcal{D}ir(\alpha)$	distribution de Dirichlet de dimension K
z	indice sur les thèmes
$z_{d,i} = k$	le $i^{\text{ème}}$ mot dans le $d^{\text{ème}}$ document est assigné au concept k

TABLE *D.1* – *Liste des symboles utilisés pour la description du modèle LDA.*

Le modèle LDA génère un nouveau document, un mot suivant l'autre, avec comme unique information les paramètres du modèle, selon le procédé suivant :

1. Pour chaque thème : choisir les mots les plus probables.
2. Pour chacun des documents composant le corpus :
   a) Déterminer la proportion de chaque thème au sein du document.
   b) Pour chaque occurrence de mot composant le vocabulaire :
      i. Choisir un thème.
      ii. Sachant le thème préalablement choisi, sélectionner le mot le plus probable sachant le mot et le thème (**retour à l'étape 1**).

Les paramètres issus du modèle LDA sont déterminés à partir d'un tirage d'un échantillon selon une distribution de Dirichlet comme défini précédemment, et une distribution multinomiale. Cette dernière (généralisation de la distribution binomiale) indique la probabilité d'observer plusieurs événements sachant le nombre de tirages et une distribution fixe sommant à 1 comme résultat. Le procédé génératif probabiliste, défini dans (Blei et Lafferty, 2009), utilise la notation définie dans le tableau D.1 :

1. Choisir la taille N du vocabulaire V selon une loi de Poisson $N \sim \text{Poisson}(\zeta)$.
2. Pour chaque thème k, tirer une distribution sur les mots des thèmes $\phi_k \sim \mathcal{D}ir(\beta)$.
3. Pour chacun des documents d :
   (a) Tirer un vecteur de distribution de proportion de thème $\theta_d \sim \mathcal{D}ir(\alpha)$.
   (b) Pour chacun occurrence des N mots i :
      i. Choisir un thème $z_{d,i} \sim \text{Multinomial}(\theta_d)$, $z_{d,i} \in \{1, \ldots, K\}$.
      ii. Tirer un mot $w_{d,i} \sim \text{Multinomial}(\phi_{z_{d,i}})$, $w_{d,i} \in \{1, \ldots, V\}$.

Une manière classique de représenter un modèle statistique, est le modèle graphique dit "plat" (Jordan et al., 2004), comme les auteurs dans (Gelman et al., 2013) ont défini le modèle hiérarchique bayésien. La figure D.4 montre LDA dans une schématisation dite "plate". Les rectangles (N, D, K) signifient une répétition des noeuds se trouvant à l'intérieur de celles-ci.

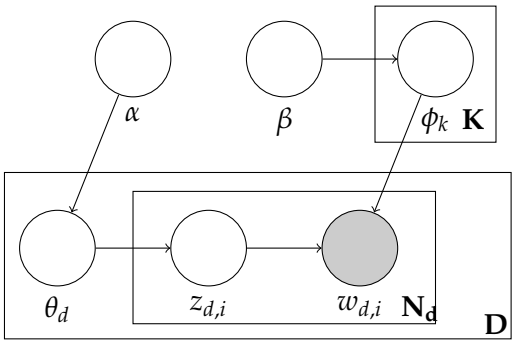


FIGURE *D.4* – *Modèle LDA.*

De la figure D.4 la distribution suivante peut alors être déduite :

$$P(w, z, \theta, \phi|\alpha, \beta) = P(\theta|\alpha)P(z|\theta)P(\phi|\beta)P(w|z, \phi). \tag{D.3}$$

Les probabilités contenues dans le modèle sont obtenues en marginalisant cette probabilité jointe. Ce modèle se caractérise donc par les quatre probabilités décrites ci-dessus. Elles ont, chacune, une signification et un rôle précis dans le modèle LDA. Nous allons les détailler dans les sections suivantes.

D.2.1 Distribution des thèmes dans un document

Cette probabilité est tirée selon une loi de Dirichlet de paramètre α (voir figure D.5) composant le vecteur de dimension K ($\alpha_k > 0$) :

$$P(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \theta_1^{\alpha_1} \ldots \theta_K^{\alpha_K} = \frac{\Gamma(\alpha.)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \tag{D.4}$$

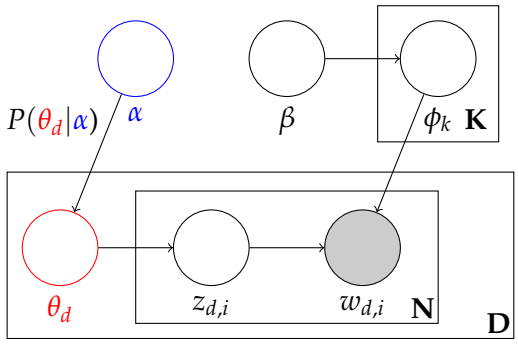


FIGURE *D.5* – *Distribution des thèmes au sein d'un document dans le modèle LDA.*

où (.) désigne le produit scalaire pour des raisons de simplification d'écriture. Cette probabilité contrôle la distribution des thèmes au sein des documents. Elle est calculée une seule fois, et assure alors une homogénéité entre les distributions des thèmes au sein des documents.

D.2.2 Attribution des mots au sein des thèmes

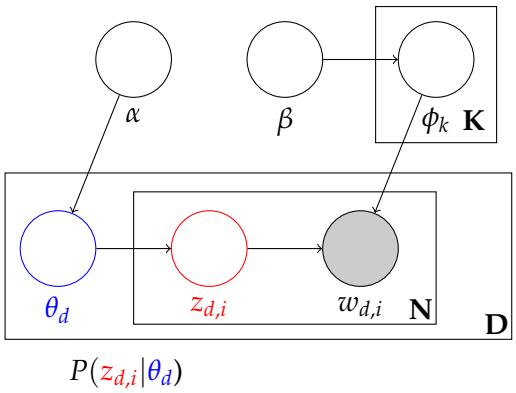


FIGURE *D.6* – *Attribution des mots au sein des thèmes dans le modèle LDA.*

Dans le corpus de documents utilisé pour l'estimation du modèle LDA, l'attribution des termes composant le vocabulaire dépend de la distribution z, qui dépend elle-même de θ (voir figure D.6). Ainsi, à chacune des occurrences des N mots composant le vocabulaire V, une valeur comprise entre $\{1, \ldots, K\}$ lui est assignée. Chaque thème k est assigné $n_{d,k}$ fois dans un même document, autant que le nombre de mots du document d assigné au thème k :

$$P(z|\theta) = \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{d,k}^{n_{d,k}} \tag{D.5}$$

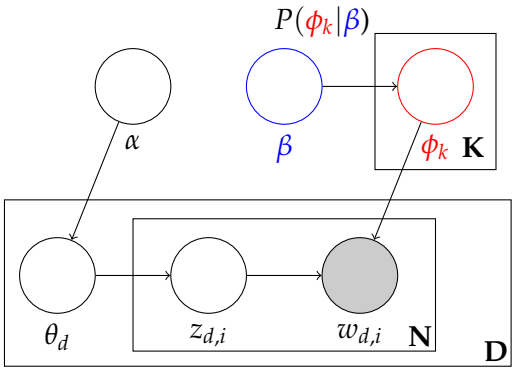


FIGURE D.7 – *Probabilité de tirage d'un mot sachant le tirage du thème k.*

D.2.3 Probabilité du choix d'un mot

La probabilité liant ϕ et β contrôle la distribution des termes composant le vocabulaire V au sein des thèmes du modèle LDA.

$$P(\phi|\beta) = \prod_{k=1}^{K} \frac{\Gamma(\beta_{k,.})}{\prod_{v=1}^{V} \Gamma(\beta_{k,v})} \phi_{k,v}^{\beta_{k,v}-1}, \tag{D.6}$$

où $\phi_{k,v}$ est la probabilité qu'un mot v soit tiré sachant que le thème choisi est le thème k.

D.2.4 Probabilité d'un corpus de documents

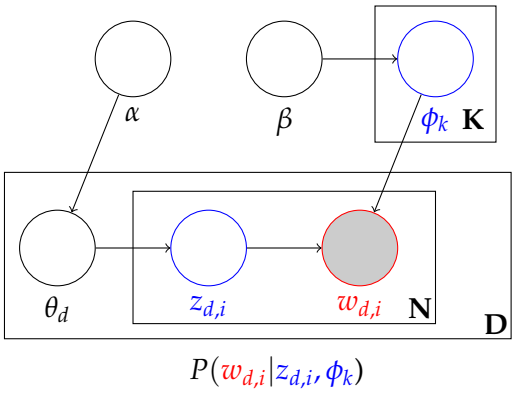


FIGURE D.8 – *Probabilité du corpus.*

Etant donné un ensemble de documents, la probabilité d'un terme w sachant les variables latentes z et ϕ (figure D.8) est donnée par :

$$P(w|z,\phi) = \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{n_{.,k,v}}, \tag{D.7}$$

où $n_{.,k,v}$ est le nombre de fois où le thème k est attribué au mot v dans l'ensemble du corpus.

D.2.5 Paramètres du modèle

La formalisation du modèle génératif LDA, définie dans l'équation D.3, est alors explicitée ainsi en utilisant les différentes écritures décrites dans les sections précédentes :

$$\begin{aligned}
P(w,z,\theta,\phi|\alpha,\beta) &= P(\theta|\alpha)P(z|\theta)P(\phi|\beta)P(w|z,\phi) \\
&= \left(\prod_{d=1}^{D}\frac{\Gamma(\alpha.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k-1}\right)\left(\prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{n_{d,k}}\right)\times \\
&\left(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k,.})}{\prod_{v=1}^{V}\Gamma(\beta_{k,v})}\prod_{v=1}^{V}\phi_{k,v}^{\beta_{k,v}-1}\right)\left(\prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{n_{.,k,v}}\right) \\
&= \left(\prod_{d=1}^{D}\frac{\Gamma(\alpha.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k+n_{d,k,.}-1}\right)\times\left(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k,.})}{\prod_{v=1}^{V}\Gamma(\beta_{k,v})}\prod_{v=1}^{V}\phi_{k,v}^{\beta_{k,v}+n_{.,k,v}-1}\right).
\end{aligned} \tag{D.8}$$

L'équation D.8 nécessite d'être marginalisée pour écrire un modèle probabiliste d'un corpus de documents connaissant les hyper-paramètres α et β et pour estimer le maximum de vraisemblance des paramètres du modèle ainsi que pour la distribution des variables latentes :

$$\begin{aligned}
P(w,z|\alpha,\beta) &= \int_\phi\int_\theta\left(\prod_{d=1}^{D}\frac{\Gamma(\alpha.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k+n-d,k,.-1}\right)\times \\
&\left(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k,.})}{\prod_{v=1}^{V}\Gamma(\beta_{k,v})}\prod_{v=1}^{V}\phi_{k,v}^{\beta_{k,v}+n_{.,k,v}-1}\right)\mathrm{d}\theta\mathrm{d}\phi \\
&= \int_\theta\left(\prod_{d=1}^{D}\frac{\Gamma(\alpha.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k+n-d,k,.-1}\right)\mathrm{d}\theta\times \\
&\int_\phi\left(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k,.})}{\prod_{v=1}^{V}\Gamma(\beta_{k,v})}\prod_{v=1}^{V}\phi_{k,v}^{\beta_{k,v}+n_{.,k,v}-1}\right)\mathrm{d}\phi
\end{aligned} \tag{D.9}$$

D.3 Espace de thèmes LDA

Le modèle génératif probabiliste LDA permet donc d'extraire, depuis un corpus de documents, la relation latente liant les occurrences des mots composant ces documents.

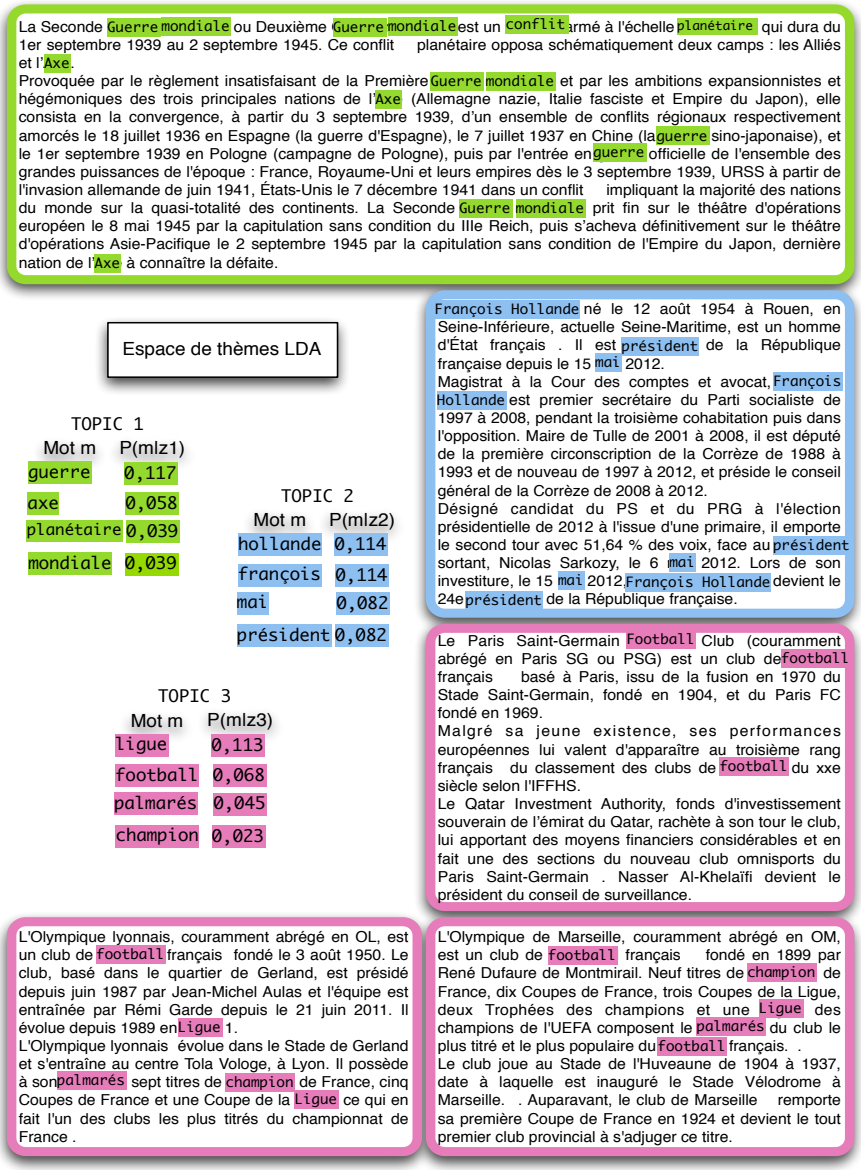


FIGURE *D.9* – *Exemple d'un espace de thèmes LDA.*

Elle est représentée par une distribution de thèmes au sein de chacun de ces documents. Dans cette section, un modèle LDA de trois thèmes est présenté (voir figure D.9).

Il est ainsi observé, dans un premier temps, que les thèmes trouvés par la méthode LDA sont proches de ceux trouvés par la méthode LSA B.4. La différence majeure est l'absence du terme *France* dans un des thèmes composant l'espace LDA, contrairement aux thèmes issus de LDA. Les thèmes LDA contiennent les termes les plus discriminants de chacun des articles (guerre, François, Hollande, football), et sont fortement représentatifs des thèmes (histoire, politique et football) composant ce corpus.

Annexe E

Système de reconnaissance automatique de la parole (SRAP) Speeral

Cette annexe décrit le système de reconnaissance de la parole Speeral (Linarès et al., 2007) dans sa version destinée au décodage des dialogues du centre d'appel de la RATP. Ce système a été développé au LIA, l'adaptation au cas d'utilisation RATP ayant été réalisée dans le cadre du projet ANR DECODA.

Speeral est un système dont les bases technologiques sont assez classiques. Les modèles acoustiques sont des HMM gauche-droite à trois états partagés. Les modèles de langage sont des n-grammes, typiquement 3 ou 4 grammes en fonction des contextes d'utilisation. Le système exécute deux passes, la seconde impliquant des modèles de langage adaptés au locuteur par MLLR.

E.1 Algorithme de recherche

Speeral utilise l'algorithme de recherche $A^\star$, relativement courant en reconnaissance de la parole. C'est un algorithme asynchrone qui développe le graphe de recherche en différents points choisis par une approximation des coûts des chemins complets. Par contre, le fait que le décodeur opère sur un treillis phonétique est moins courant. Cette stratégie permet de réaliser très tôt des coupures qui permettent d'accélérer le système. Une étude complète du système et de ses performance peut être trouvée dans (Linarès et al., 2007).

E.2 Paramètres et Modèles acoustiques

Le système utilise une paramétrisation PLP avec une normalisation (cepstres centrés-réduits) sur une fenêtre glissante de 30ms déplacée de 10ms. Les modèles reposent sur une base classique HMM/GMM, avec des états partagés par la méthode des arbres de décision. Le système grand vocabulaire "généraliste" qui a servi de base au système DECODA contient 230K gaussiennes et 3 600 états émetteurs pour 10 000 HMMs. Les GMMs ont été appris par un algorithme EM. Les modèles acoustiques de DECODA sont issus de la version grand vocabulaire de Speeral et ont été appris sur une version filtrée à 8Khz de 250 heures de parole annotée (issu des corpus ESTER 1 et 2 et du corpus EPAC). Les modèles sont ensuite adaptés, par MAP, sur les 30 heures du corpus DECODA.

E.3 Modèle de langage

Le modèle de langage a été complètement ré-appris sur les données transcrites de DECODA. Le vocabulaire est réduit à 27K mots (contre 88k pour le système standard) et les modèles sont réduits aux 3-grammes (4-grammes pour la version générique). Comme pour la version générique, les modèles sont estimés par le *toolkit* SRI.

Le vocabulaire du corpus DECODA contient 5 782 mots et le taux d'erreur-mot (TEM) initial atteint 45,8 % sur le corpus d'apprentissage et 58,0 % sur le corpus de test. Ces TEM élevés sont principalement dus à la présence de nombreuses disfluences, et à des conditions acoustiques bruitées, quand, par exemple, les utilisateurs appellent à partir de gares avec un téléphone portable. Une liste de rejet (*stop-list*) de 126 mots [1] a été utilisée, ce qui donne au final un TEM de 33,8 % (apprentissage), et 49,5 % sur l'ensemble (test).

1. **http://code.google.com/p/stop-words/**

Annexe F

Analyse Factorielle

L'analyse factorielle ou *Factor analysis* (FA) trouve ses origines dans le domaine de recherche lié à l'analyse statistique. Cette méthode, appelée analyse en composantes principales ou *Principal Component Analysis* (ACP), a pour vocation première de réduire un espace de représentation tout en maximisant la variance des projetés (voir section 2.2.6.3). Ce modèle présente néanmoins certains inconvénients tels l'absence d'un modèle génératif des données et d'une densité de probabilité associée (Jolliffe, 1986).

Une version probabiliste de l'ACP est alors proposée par (Tipping et Bishop, 1999) et appelée analyse en composantes principales probabiliste (ACPP). Cette méthode utilise le critère du maximum de vraisemblance sur un modèle à variables latentes, pour déterminer les axes de projection. Les auteurs dans (Kenny et al., 2005) proposent une extension de l'ACPP dans le domaine de la reconnaissance de la parole. Les travaux proposés dans ce manuscrit s'appuient essentiellement sur cette extension. Cette méthode considère les vecteurs de représentation dans un espace formé par la concaténation des moyennes des gaussiennes issues de mixtures de gaussiennes, modélisant les différentes parties de l'espace acoustique. La section suivante propose une description de l'analyse factorielle.

F.1 Analyse factorielle

L'analyse factorielle est une généralisation de l'ACP probabiliste (ACPP). De la même manière que dans l'ACPP, le vecteur observé y de dimension d est généré à partir du vecteur caché t de dimension r suivant l'expression :

$$y = At + \mu + \varepsilon \tag{F.1}$$

La différence avec l'ACPP se situe dans l'hypothèse faite concernant la distribution du bruit ε. Dans le cas de l'analyse factorielle, le bruit est supposé être gaussien de

moyenne nulle et de matrice de covariance Σ diagonale. Les éléments de Σ sont différents d'une composante à l'autre, contrairement à l'ACPP.

Les modèles fondés sur les mixtures de gaussiennes ou *Gaussian Mixture Model* (GMM) sont généralement utilisés pour approximer des densités de probabilité. Les GMM sont estimés, généralement, en utilisant le critère du maximum de vraisemblance, sur des données observées (appelées *réalisations*). Les données observées englobent différentes caractéristiques, parmi lesquelles se trouve la caractéristique que nous souhaitons modéliser. Par exemple, les trames d'une session de parole contiennent des informations sur le locuteur, la canal, l'état émotionnel du locuteur, ou encore le contenu phonétique.

Lorsque nous nous intéressons aux informations sur le locuteur, le super-vecteur associé au GMM correspondant aux données observées sera noté m_l. Si, par contre, nous nous intéressons à l'information canal, le super-vecteur sera noté m_c. D'une manière générale, nous noterons le super-vecteur m_e, avec e étant une caractéristique quelconque : locuteur, canal, ...

Les moyennes d'un GMM peuvent être vues comme des variables aléatoires dont les caractéristiques statistiques peuvent être obtenues en utilisant l'analyse factorielle. Une manière de faire est d'appliquer l'équation F.1 à chacune des moyennes des gaussiennes du mélange :

$$m_e^i = m^i + A^i t^i + \varepsilon \tag{F.2}$$

avec m_e^i le vecteur moyenne de la gaussienne i pour la caractéristique e. L'inconvénient de ce modèle est que les différentes composantes de la mixture de gaussiennes sont traitées indépendamment les unes des autres. Ce qui ne permet pas de mettre en évidence la corrélation pouvant exister entre les différentes régions (gaussiennes) de l'espace des observations. La solution proposée par (Kenny et al., 2005) est d'appliquer le modèle d'analyse factorielle à une nouvelle variable aléatoire obtenue par la concaténation des moyennes du GMM. Cette nouvelle variable aléatoire est appelée *super-vecteur*. Dans ce cadre, tous les GMM sont obtenus à partir d'un GMM global, estimé sur une grande quantité de données et incluant le plus de variabilités possibles. Ce GMM est appelé modèle du monde, ou UBM (*Universal Background Model*). Soit m le super-vecteur associé à l'UBM. Le super-vecteur observation m_e est obtenu comme suit :

$$m_e = m + \mathbf{U}\mathbf{x}_e + \varepsilon \tag{F.3}$$

Il est important de souligner le fait que dans ce modèle, le vecteur x_e est commun à toutes les gaussiennes. Cette caractéristique permet d'exploiter la redondance pouvant exister entre les différentes gaussiennes. Ce vecteur x_e est une représentation compacte de m_e.

F.2 Estimation des paramètres du modèle d'analyse factorielle

Dans le modèle de l'équation F.1, la variable aléatoire en question est le super-vecteur. Il faut donc estimer les super-vecteurs à partir des observations (trames dans le cas de la parole). Ensuite, ces super-vecteurs peuvent être utilisés pour estimer la matrice de projection ($\mathbf{U}$) et les différents vecteurs projetés ($\mathbf{x}_e$). Cependant, il est possible d'estimer ces paramètres directement à partir des observations, sans passer par les super-vecteurs. Cette estimation peut être réalisée en utilisant le critère du maximum de vraisemblance, que nous décrivons dans le reste de cette section.

Dans la suite, nous décrivons le procédé d'estimation de la matrice $\mathbf{U}$ et des vecteurs $\mathbf{x}_e$. Avant tout, nous présentons les notations nécessaires au développement de la stratégie d'estimation :

— m_e : super-vecteur associé à la caractéristique e.
— m : super-vecteur du modèle du monde (UBM).
— Σ_g : matrice de covariance de la gaussienne g.
— $\chi(e)$: données d'apprentissage de la caractéristique e.
— $P_{GMM}(\chi(e)|m,\Sigma)$: vraisemblance de $\chi(e)$ étant donné le GMM spécifique au super-vecteur m et de super-matrice de covariance Σ.
— $N_g(e)$: nombre de vecteurs acoustiques d'un enregistrement e appartenant à la gaussienne g.
— G : nombre des gaussiennes dans le GMM-UBM.
— F : taille des vecteurs acoustiques.
— Pour chaque gaussienne g, nous calculons :

$$S_{X,g}(e) = \sum_t (X_t - \mu_g) \qquad ; \qquad S_{X^tX,g}(e) = \sum_t (X_t - \mu_g)^t (X_t - \mu_g) \tag{F.4}$$

où $\sum_t$ est la somme sur tous les vecteurs acoustiques d'un enregistrement e appartenant à la gaussienne g et μ_g est la moyenne de la gaussienne g du GMM du modèle du monde (UBM).

Pour chaque enregistrement e, $logP_{\mathbf{U},\Sigma}(\chi(e)|\mathbf{x}_e)$ est la vraisemblance conditionnelle de $\chi(e)$ étant donné x_e et de paramètres $(\mathbf{U},\Sigma)$:

$$logP_{\mathbf{U},\Sigma}(\chi(e)|\mathbf{x}_e) = G_\Sigma(e) + H_{\mathbf{U},\Sigma}(e,\mathbf{x}_e) \tag{F.5}$$

avec $G_\Sigma(e)$ et $H_{\mathbf{U},\Sigma}(e,\mathbf{x}_e)$ obtenus par les équations suivantes :

$$
\begin{aligned}
G_\Sigma(e) &= \sum_{g=1}^{G}(N_g(e)log(\frac{1}{(2\pi)^{F/2}|\Sigma_g|^{1/2}}) - \frac{1}{2}tr(\Sigma_g^{-1}S_{X^tX,g}(e)) \\
H_{\mathbf{U},\Sigma}(e,\mathbf{x}_e) &= {}^t\mathbf{x}_e^t\mathbf{U}\Sigma^{-1}S_X(e) - \frac{1}{2}{}^t\mathbf{x}^t\mathbf{U}.N(e)\Sigma^-1\mathbf{U}\mathbf{x}_e
\end{aligned}
$$

Dans la phase d'apprentissage, pour chaque enregistrement e à tester, la distribution *a posteriori* de $\mathbf{x}_e$ sachant $\chi(e)$ et les paramètres $\mathbf{U}$ et Σ suivent une loi gaussienne de moyenne $l^{-1}(e)^t\mathbf{U}\Sigma^{-1}S_X(e)$ et de matrice de covariance $l^{-1}(e)$.

Pour chaque enregistrement e, soit $G_\Sigma(e)$ le *log* de la fonction de vraisemblance gaussienne donnée par :

$$
\sum_{g=1}^{G}[N_g(e)log(\frac{1}{(2\pi)^{F/2}|\Sigma_g|^{1/2}}) - \frac{1}{2}tr(\Sigma_g^{-1}S_{X^tX,g})]
$$

alors :

$$
log(P_{\mathbf{U},\Sigma}(\chi(e))) = G_\Sigma(e) - \frac{1}{2}log|l(e)| + \frac{1}{2}{}^t(\hat{m}_s - m)\Sigma^{-1}S_X(e)
$$

Si pour chaque enregistrement e, la moyenne et la covariance de m_e *a posteriori* sont notées par $\hat{m}_e$ et $\hat{B}(e)$:

$$
\hat{m}_e = m + \mathbf{U}.l^{-1}.{}^t\mathbf{U}\Sigma^{-1}S_X(e) \tag{F.6}
$$

$$
\hat{B}(e) = \mathbf{U}.l^{-1}.{}^t\mathbf{U} \tag{F.7}
$$

donc :

$$
exp(\frac{1}{2}.{}^tS_X\Sigma^{-1}\mathbf{U}.l^{-1}.{}^t\mathbf{U}\Sigma^{-1}S_X(e)) = exp(\frac{1}{2}.{}^t(\hat{m}_e - m)\Sigma^{-1}S_X(e))
$$

d'où :

$$
logP_{\mathbf{U},\Sigma}(\chi(e)) = G_\Sigma(e) - \frac{1}{2}log|l(e)| + \frac{1}{2}{}^t(\hat{m}_e - m)\Sigma^{-1}S_X(e)
$$

Algorithme d'apprentissage EM :

Supposons que les paramètres initiaux $(\mathbf{U}_0, \Sigma_0)$ ont été estimés. Pour chaque enregistrement e, soient $E[\mathbf{x}_e]$ et $E[\mathbf{x}_e^t\mathbf{x}_e]$ les premiers et seconds moments de $\mathbf{x}_e$ calculés avec les équations suivantes :

$$
E[\mathbf{x}_e] = l^{-1}.{}^t\mathbf{U}\Sigma^{-1}S_X(e) \qquad ; \qquad E[\mathbf{x}_e.{}^t\mathbf{x}_e] = E[\mathbf{x}_e]E[{}^t\mathbf{x}_e] + l^{-1} \tag{F.8}
$$

Soit $(\mathbf{U}, \Sigma)$ la nouvelle estimation des paramètres du modèle définie comme suit :

$$\sum_s N(e)\mathbf{U}.E[\mathbf{x}_e^t \mathbf{x}_e] = \sum_s S_X(e)E[^t\mathbf{x}_e] \tag{F.9}$$

et : $\forall g \in 1, ..., G$:

$$\Sigma_g = \frac{1}{n_g}\sum_e S_{X^t X,g}(e) - M_g \tag{F.10}$$

où

$$logP_{\mathbf{U},\Sigma}(\chi(e)) = G_\Sigma(e) - \frac{1}{2}log|l(e)| + \frac{1}{2}^t(\hat{m}_s - m)\Sigma^{-1}S_X(e)$$

avec $n_g = \sum_s N_g(e)$ et M_g représentant le $g^{ième}$ bloc de la matrice diagonale par bloc $CF \times CF$:

$$\frac{1}{2}\sum_e (S_X(e)E[^t\mathbf{x}(e)]^t\mathbf{U} + \mathbf{U}.E[\mathbf{x}_e)]^t S_X(e))$$

Puis :

$$\sum_e P_{\mathbf{U},\Sigma}(\chi(e)) \geq \sum_e P_{\mathbf{U}_0,\Sigma_0}(\chi(e))$$

L'équation F.9 est le système linéaire d'équations pour notre problème linéaire de régression. Pour le résoudre, remarquons que la matrice $N(e)$ est diagonale, ce qui rend égales les $i^{ième}$ lignes de chaque coté de l'égalité (i $\in \{1, ..., CF\}$). Ceci donne :

$$\mathbf{U}^i\sum_s N^i(e)E[\mathbf{x}_e^t\mathbf{x}_e] = \sum_s S_X^i(e)E[^t\mathbf{x}_e]$$

où X^i est la $i^{ième}$ ligne de la matrice X. Remarquons aussi que si l'on écrit i sous la forme $(g-1)F + f$ où $1 \leq g \leq G$ et $1 \leq f \leq F$, alors $N^i(e) = N_g(e)$, et nous obtenons :

$$\mathbf{U}^i\sum_e N_g(e)E[\mathbf{x}_e^t\mathbf{x}_e)] = \sum_e S_X^i(e)E[^t\mathbf{x}_e] \tag{F.11}$$

Grâce à l'équation F.11, la matrice $\mathbf{U}$ peut être estimée ligne par ligne. L'algorithme 4 présente la stratégie adoptée pour estimer la matrice $\mathbf{U}$ et les vecteurs x_e avec les développements ci-dessus.

Une démonstration détaillée est présentée dans l'annexe A.

Algorithm 4: Algorithme d'estimation de la matrice **U** et du vecteur **x**.

Pour chaque enregistrement e : $x_{(e)} \leftarrow 0$, $\mathbf{U} \leftarrow random$;
Estimation de : $S_{X,g}(e), S_{X^tX,g}(e)$; (eq : F.4)
Estimation de : $E[\mathbf{x}_e], E[\mathbf{x}_e.{}^t\mathbf{x}_e]$; (eq : F.8)
pour $i = 1$ *to nb_iterations* **faire**
  pour *tous les enregistrements s* **faire**
    Estimation $\hat{m}_e$;(eq : F.6)
    Estimation $\hat{B}(e)$;(eq : F.7)
  fin
  Estimation de la matrice **U** ;(eq : F.11)
fin

F.3 Analyse factorielle pour la vérification de locuteur

Dans cette thèse, nous utilisons l'approche de l'analyse factorielle dans la modélisation acoustique pour la reconnaissance automatique de la parole. Le modèle d'analyse factorielle que nous utilisons est inspiré de la modélisation par analyse factorielle dans le domaine de la vérification de locuteur (Kenny et al., 2005). Dans le Système de Vérification du Locuteur (SVL), cette modélisation a permis d'améliorer la robustesse face à la variabilité session qui représente une cause majeure de dégradation des performances des systèmes. Dans cette section, nous présentons brièvement le modèle d'analyse factorielle appliqué à la vérification de locuteurs.

Dans le cadre d'un SVL, le modèle du monde (GMM-UBM) représente la génération de vecteurs cepstraux provenant d'une multitude de locuteurs et de sessions. L'estimation des paramètres de ce modèle est réalisée en utilisant l'algorithme EM (Expectation-Maximisation, voir section C.3). Le GMM d'un locuteur donné est obtenu à partir de l'UBM en ré-estimant les moyennes. Les poids et les variances restent inchangés. Cette adaptation des moyennes est réalisée en utilisant l'approche d'analyse factorielle. Pour prendre en compte la variabilité liée au canal d'enregistrement, le modèle de l'équation F.3 est étendu comme suit :

$$m_{h,l} = m + \mathbf{D}\mathbf{y}_l + \mathbf{U}\mathbf{x}_{h,l} \tag{F.12}$$

Dans ce modèle, le terme $m + Dy_l$ modélise cette fois la part propre au locuteur. **D** est une matrice diagonale de taille $GF \times GF$ et y_l est un vecteur de dimension GF estimé sur les données du locuteur l. **D** satisfait l'équation $\mathbf{I} = \tau\mathbf{D}^t\Sigma^{-1}\mathbf{D}$ où τ est un facteur appelé *relevance factor*. Le facteur $\mathbf{U}\mathbf{x}_{h,l}$ est la composante introduite par l'effet de la session (canal). Les vecteurs colonnes de la matrice **U** ($FG \times r$) représentent une base du sous-espace dans lequel évolue la variabilité session. $\mathbf{x}_{h,l}$ est un vecteur de dimension r contenant les composantes relatives à la session dans ce sous-espace.

La matrice **U** et le vecteur $\mathbf{x}_e$ sont estimés selon la description de la section F.2. Mais, dans ce modèle, les statistiques de l'équation F.4 sont calculées à nouveau comme suit :

$$S_{X,g}(e) = \sum_t (X_t - m_g - \mathbf{Dy})$$

$$S_{X^t X,g}(e) = \sum_t (X_t - \mathbf{m}_g - \mathbf{Dy})^t (X_t - \mathbf{m}_g - \mathbf{Dy})$$

où $\sum_t$ est la somme sur tous les vecteurs acoustiques du locuteur l appartenant à la gaussienne g et $\mathbf{m}_g$ est la moyenne de la gaussienne g du GMM-UBM. Les sommes se font sur tous les vecteurs acoustiques du locuteur l appartenant à la gaussienne g, et $\mathbf{m}_g$ est la moyenne de la gaussienne g du GMM du monde.
Dans ce modèle, le vecteur y_l est estimé sur les données du locuteur l en s'appuyant sur l'équation suivante :

$$\{y_l\}_g = \frac{\tau}{(\tau + N_l(g))} . \mathbf{D}_g . \Sigma^{-1} . [\mathbf{X}_g^l - \sum_{h \in l} . \{\mathbf{m} + \mathbf{U}\mathbf{x}_{(h,l)}\}[g]] \tag{F.13}$$

avec $D_g = \frac{\Sigma_g^{1/2}}{\sqrt{\tau}}$.

F.4 Espace de variabilité totale

L'analyse factorielle, présentée plus haut, est devenue une méthode de référence pour la réduction de dimension en vérification du locuteur (Dehak et al., 2011). Ainsi, les auteurs présentent pour la première fois une version compacte $\mathbf{w}$, appelée *i*-vecteur. Le procédé d'extraction peut être vu comme un processus de compression probabiliste réduisant la dimensionnalité du super-vecteur du segment de parole en suivant un modèle linéaire gaussien. Le super-vecteur $\mathbf{m}_s$ d'un segment de parole composé des concaténations des moyennes de l'UBM-GMM est projeté dans un espace de dimension réduit, appelé espace de variabilité totale ou *Total variability space*, avec :

$$\mathbf{m}_s = m + \mathbf{T}\mathbf{x}_s \, , \tag{F.14}$$

où m est le super-vecteur représentant les moyennes de l'UBM-GMM [1]. $\mathbf{T}$ est une matrice de faible dimension $(MD \times R)$ avec M le nombre de gaussiennes composant l'UBM, et D la taille du vecteur cepstral représentant une base de l'espace de variabilité totale. $\mathbf{T}$ est appelée matrice de variabilité totale ; les composantes de $\mathbf{x}_s$ sont les facteurs représentant les coordonnées de l'enregistrement audio dans l'espace de variabilité totale appelé *i*-vector (*i* pour *i*dentification).

1. L'UBM est un GMM représentant toutes les observations.

Liste des illustrations

Liste des tableaux

Bibliographie

(Abdi et Williams, 2010) H. Abdi & L. J. Williams, 2010. Principal component analysis. *Wiley Interdisciplinary Reviews : Computational Statistics* *2*(4), 433–459.

(Akita et Kawahara, 2003) Y. Akita & T. Kawahara, 2003. Unsupervised speaker indexing using anchor models and automatic transcription of discussions. Dans les actes de *INTERSPEECH*.

(Akita et Kawahara, 2004) Y. Akita & T. Kawahara, 2004. Language model adaptation based on plsa of topics and speakers. Dans les actes de *INTERSPEECH*.

(Alexiadis et Sergiadis, 2009) D. S. Alexiadis & G. D. Sergiadis, 2009. Estimation of motions in color image sequences using hypercomplex fourier transforms. *Image Processing, IEEE Transactions on* *18*(1), 168–187.

(Allan et al., 1998a) J. Allan, J. Carbonell, G. Doddington, J. Yamron, & Y. Yang, 1998a. Topic detection and tracking pilot study final report.

(Allan et al., 1998b) J. Allan, R. Papka, & V. Lavrenko, 1998b. On-line new event detection and tracking. Dans les actes de *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 37–45. ACM.

(Arena et al., 1994) P. Arena, L. Fortuna, L. Occhipinti, & M. G. Xibilia, 1994. Neural networks for quaternion-valued function approximation. Dans les actes de *Circuits and Systems, 1994. ISCAS'94., 1994 IEEE International Symposium on*, Volume 6, 307–310. IEEE.

(Aronowitz, 2007) H. Aronowitz, 2007. Trainable speaker diarization. Dans les actes de *INTERSPEECH*, 1861–1864.

(Aronowitz, 2010) H. Aronowitz, 2010. Unsupervised compensation of intra-session intra-speaker variability for speaker diarization. Dans les actes de *Odyssey*, 25.

(Arun et al., 2010) R. Arun, V. Suresh, C. Veni Madhavan, & M. Narasimha Murthy, 2010. On finding the natural number of topics with latent dirichlet allocation : Some observations. Dans les actes de *Advances in Knowledge Discovery and Data Mining*, 391–402. Springer.

(Aspragathos et Dimitros, 1998) N. A. Aspragathos & J. K. Dimitros, 1998. A comparative study of three methods for robot kinematics. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on* *28*(2), 135–145.

(Assange, 2013) J. Assange, 2013. La vraie guerre, c'est la guerre de l'information. http://www.courrierinternational.com/article/2013/07/10/julian-assange-la-vraie-guerre-c-est-la-guerre-de-l-information?page=all.

(Assefa et al., 2010) D. Assefa, L. Mansinha, K. F. Tiampo, H. Rasmussen, & K. Abdella, 2010. Local quaternion fourier transform and color image texture analysis. *Signal Processing* *90*(6), 1825–1835.

(Asuncion et Newman, 2007) A. Asuncion & D. Newman, 2007. Uci machine learning repository.

(Asuncion et al., 2009) A. Asuncion, M. Welling, P. Smyth, & Y. W. Teh, 2009. On smoothing and inference for topic models. Dans les actes de *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 27–34. AUAI Press.

(Azzopardi et al., 2004) L. Azzopardi, M. Girolami, & C. Van Rijsbergen, 2004. Topic based language models for ad hoc information retrieval. Dans les actes de *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, Volume 4, 3281–3286. IEEE.

(Baeza-Yates et al., 1999) R. Baeza-Yates, B. Ribeiro-Neto, et al., 1999. *Modern information retrieval*, Volume 463. ACM press New York.

(Bahri et Hitzer, 2007) M. Bahri & E. S. Hitzer, 2007. Clifford algebra cl3, 0-valued wavelet transformation, clifford wavelet uncertainty inequality and clifford gabor wavelets. *International Journal of Wavelets, Multiresolution and Information Processing* *5*(06), 997–1019.

(Balakrishnan et Nevzorov, 2004) N. Balakrishnan & V. B. Nevzorov, 2004. *A primer on statistical distributions*. John Wiley & Sons.

(Bartlett et Shawe-Taylor, 1999) P. Bartlett & J. Shawe-Taylor, 1999. Generalization performance of support vector machines and other pattern classifiers. *Advances in Kernel Methods-Support Vector Learning*, 43–54.

(Bass, 2004) F. Bass, 2004. Comments on "a new product growth for model consumer durables the bass model". *Management science* *50*(12 supplement), 1833–1840.

(Batard et al., 2010) T. Batard, M. Berthier, & C. Saint-Jean, 2010. Clifford–fourier transform for color image processing. Dans les actes de *Geometric Algebra Computing*, 135–162. Springer.

(Bayro-Corrochano, 2006) E. Bayro-Corrochano, 2006. The theory and use of the quaternion wavelet transform. *Journal of Mathematical Imaging and Vision* *24*(1), 19–35.

(Bayro-Corrochano et Arana-Daniel, 2010) E. J. Bayro-Corrochano & N. Arana-Daniel, 2010. Clifford support vector machines for classification, regression, and recurrence. *Neural Networks, IEEE Transactions on 21*(11), 1731–1746.

(Bechet et al., 2012) F. Bechet, B. Maza, N. Bigouroux, T. Bazillon, M. El-Beze, R. De Mori, & E. Arbillot, 2012. Decoda : a call-centre human-human spoken conversation corpus. LREC'12.

(Beigi et Maes, 1998) H. S. Beigi & S. Maes, 1998. Speaker, channel and environment change detection. Dans les actes de *Proc. of the World Congress on Automation*.

(Bellegarda, 1997) J. Bellegarda, 1997. A latent semantic analysis framework for large-span language modeling. Dans les actes de *Fifth European Conference on Speech Communication and Technology*.

(Bellegarda, 2000) J. Bellegarda, 2000. Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE 88*(8), 1279–1296.

(Bengio, 2009) Y. Bengio, 2009. Learning deep architectures for ai. *Foundations and trends® in Machine Learning 2*(1), 1–127.

(Bengio et al., 2006) Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, & J.-L. Gauvain, 2006. Neural probabilistic language models. Dans les actes de *Innovations in Machine Learning*, 137–186. Springer.

(Bentley, 1975) J. Bentley, 1975. Multidimensional binary search trees used for associative searching. *Communications of the ACM 18*(9), 509–517.

(Berry et al., 1995) M. Berry, S. Dumais, & G. O'Brien, 1995. Using linear algebra for intelligent information retrieval. *SIAM review 37*(4), 573–595.

(Berry, 1992) M. W. Berry, 1992. Large-scale sparse singular value computations. *International Journal of Supercomputer Applications 6*(1), 13–49.

(Berry et Kogan, 2010) M. W. Berry & J. Kogan, 2010. *Text mining : applications and theory*. John Wiley & Sons.

(Blei et al., 2003) D. Blei, A. Ng, & M. Jordan, 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research 3*, 993–1022.

(Blei et Lafferty, 2006a) D. M. Blei & J. Lafferty, 2006a. Correlated topic models. *Advances in neural information processing systems 18*, 147.

(Blei et Lafferty, 2006b) D. M. Blei & J. D. Lafferty, 2006b. Dynamic topic models. Dans les actes de *Proceedings of the 23rd international conference on Machine learning*, 113–120. ACM.

(Blei et Lafferty, 2009) D. M. Blei & J. D. Lafferty, 2009. Topic models. *Text mining : classification, clustering, and applications 10*, 71.

(Blei et al., 2007) D. M. Blei, J. D. Lafferty, et al., 2007. A correlated topic model of science. *The Annals of Applied Statistics 1*(1), 17–35.

(Blei et McAuliffe, 2007) D. M. Blei & J. D. McAuliffe, 2007. Supervised topic models. Dans les actes de *NIPS*, Volume 7, 121–128.

(Boser et al., 1992) B. Boser, I. Guyon, & V. Vapnik, 1992. A training algorithm for optimal margin classifiers. Dans les actes de *5th annual workshop on Computational learning theory*, 144–152.

(Bouallegue et al., 2012) M. Bouallegue, E. Ferreira, D. Matrouf, G. Linares, M. Goudi, & P. Nocera, 2012. Lia, univ. of avignon, avignon, france. Dans les actes de *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 330–335. IEEE.

(Bouallegue et al., 2011) M. Bouallegue, D. Matrouf, & G. Linares, 2011. A simplified subspace gaussian mixture to compact acoustic models for speech recognition. Dans les actes de *ICASSP*, 4896–4899.

(Bouallegue et al., 2012) M. Bouallegue, M. Rouvier, D. Matrouf, & G. Linares, 2012. Noise compensation for speech recognition using subspace gaussian mixture models. *INTERSPEECH*.

(Bousquet et al., 2011) P.-M. Bousquet, D. Matrouf, & J.-F. Bonastre, 2011. Intersession compensation and scoring methods in the i-vectors space for speaker recognition. Dans les actes de *INTERSPEECH*, 485–488.

(Brown et al., 1992) P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, & J. C. Lai, 1992. Class-based n-gram models of natural language. *Computational linguistics 18*(4), 467–479.

(Buckley et al., 1993) C. Buckley, G. Salton, & J. Allan, 1993. Automatic retrieval with locality information using smart. Dans les actes de *Proceedings of the First Text REtrieval Conference TREC-1*, 59–72.

(Bujack et al., 2014) R. Bujack, G. Scheuermann, & E. Hitzer, 2014. Detection of outer rotations on 3d-vector fields with iterative geometric correlation and its efficiency. *Advances in Applied Clifford Algebras 24*(2), 403–421.

(Bülow, 1999) T. Bülow, 1999. *Hypercomplex spectral signal representations for image processing and analysis*. Thèse de Doctorat, PhD thesis, University of Kiel, Advisor-Gerald Sommer.

(Buntine, 2009) W. Buntine, 2009. Estimating likelihoods for topic models. Dans les actes de *Advances in Machine Learning*, 51–64. Springer.

(Campbell et Higgins, 1994) J. Campbell & A. Higgins, 1994. Yoho speaker verification corpus ldc94s16. *Available at the LDC website : http ://www. ldc. upenn. edu*.

(Campbell, 2002) W. M. Campbell, 2002. Generalized linear discriminant sequence kernels for speaker recognition. Dans les actes de *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, Volume 1, I–161. IEEE.

(Campbell et al., 2006) W. M. Campbell, D. E. Sturim, D. A. Reynolds, & A. Solomonoff, 2006. Svm based speaker verification using a gmm supervector kernel and nap variability compensation. Dans les actes de *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Volume 1, I–I. IEEE.

(Cao et al., 2009) J. Cao, T. Xia, J. Li, Y. Zhang, & S. Tang, 2009. A density-based method for adaptive lda model selection. *Neurocomputing 72*(7), 1775–1781.

(Cao et Fei-Fei, 2007) L. Cao & L. Fei-Fei, 2007. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. Dans les actes de *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 1–8. IEEE.

(Carpenter, 2010) B. Carpenter, 2010. Integrating out multinomial parameters in latent dirichlet allocation and naive bayes for collapsed gibbs sampling. Rapport technique, Technical report, LingPipe.

(Cauchy, 1829) A. Cauchy, 1829. Sur l'équation à l'aide de laquelle on détermine les inégalités séculaires des mouvements des planètes. *Oeuvres Complètes (IIÃ¨me Série) 9*.

(Chan et al., 2004) W. L. Chan, H. Choi, & R. G. Baraniuk, 2004. Directional hypercomplex wavelets for multidimensional signal analysis and processing. Dans les actes de *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, Volume 3, iii–996. IEEE.

(Chang et Lin, 2011) C.-C. Chang & C.-J. Lin, 2011. Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*(3), 27.

(Chen et al., 2010) B. Chen, H. Shu, H. Zhang, G. Chen, & L. Luo, 2010. Color image analysis by quaternion zernike moments. Dans les actes de *Pattern Recognition (ICPR), 2010 20th International Conference on*, 625–628. IEEE.

(Chen et al., 2012) B. Chen, H. Shu, H. Zhang, G. Chen, C. Toumoulin, J.-L. Dillenseger, & L. M. Luo, 2012. Quaternion zernike moments and their invariants for color image analysis and object recognition. *Signal Processing 92*(2), 308–318.

(Chen et al., 2011) M. Chen, X. Jin, & D. Shen, 2011. Short text classification improved by learning multi-granularity topics. Dans les actes de *IJCAI*, 1776–1781.

(Chen et Gopalakrishnan, 1998) S. Chen & P. Gopalakrishnan, 1998. Speaker, environment and channel change detection and clustering via the bayesian information criterion. Dans les actes de *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 127–132.

(Chib, 1995) S. Chib, 1995. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association 90*(432), 1313–1321.

(Chien et Chueh, 2008) J.-T. Chien & C.-H. Chueh, 2008. Latent dirichlet language model for speech recognition. Dans les actes de *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, 201–204. IEEE.

(Chien et Chueh, 2011) J.-T. Chien & C.-H. Chueh, 2011. Dirichlet class language models for speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on 19*(3), 482–495.

(Chien et Wu, 2008) J.-T. Chien & M.-S. Wu, 2008. Adaptive bayesian latent semantic analysis. *Audio, Speech, and Language Processing, IEEE Transactions on 16*(1), 198–207.

(Chikhi et al., 2010) N. F. Chikhi, B. Rothenburger, & N. Aussenac-Gilles, 2010. Une approche probabiliste pour l'identification de structures de communautés. Dans les actes de *EGC*, 175–180.

(Choudhury et al., 2007) M. Choudhury, R. Saraf, V. Jain, S. Sarkar, & A. Basu, 2007. Investigation and modeling of the structure of texting language. Dans les actes de *IJCAI-Workshop on Analytics for Noisy Unstructured Text Data*, 63–70.

(Chueh et Chien, 2010) C.-H. Chueh & J.-T. Chien, 2010. Topic cache language model for speech recognition. Dans les actes de *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 5194–5197. IEEE.

(Claveau, 2012) V. Claveau, 2012. Vectorisation, okapi et calcul de similarité pour le tal : pour oublier enfin le tf-idf. Dans les actes de *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*.

(Clifford, 1873) W. Clifford, 1873. Preliminary sketch of bi-quaternions. *Proceedings of the London Mathematical Society* (4), 381–395.

(Cohen, 1996) W. W. Cohen, 1996. Learning rules that classify e-mail. Dans les actes de *AAAI spring symposium on machine learning in information access*, Volume 18, 25. California.

(Collobert et al., 2011) R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, & P. Kuksa, 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research 12*, 2493–2537.

(Cortes et Vapnik, 1995) C. Cortes & V. Vapnik, 1995. Support-vector networks. *Machine learning 20*(3), 273–297.

(Cullum et Willoughby, 1985) J. K. Cullum & R. A. Willoughby, 1985. *Lanczos Algorithms for Large Symmetric Eigenvalue Computations : Vol. 1 : Theory*, Volume 1.

(Daniilidis, 1999) K. Daniilidis, 1999. Hand-eye calibration using dual quaternions. *The International Journal of Robotics Research 18*(3), 286–298.

(Deerwester et al., 1990) S. Deerwester, S. Dumais, G. Furnas, T. Landauer, & R. Harshman, 1990. Indexing by latent semantic analysis. *Journal of the American society for information science 41*(6), 391–407.

(Dehak et Chollet, 2006) N. Dehak & G. Chollet, 2006. Support vector gmms for speaker verification. Dans les actes de *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006 : The*, 1–4. IEEE.

(Dehak et al., 2007) N. Dehak, P. Dumouchel, & P. Kenny, 2007. Modeling prosodic features with joint factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on 15*(7), 2095–2103.

(Dehak et al., 2011) N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, & P. Ouellet, 2011. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on 19*(4), 788–798.

(Del Moral et al., 2006) P. Del Moral, A. Doucet, & A. Jasra, 2006. Sequential monte carlo samplers. *Journal of the Royal Statistical Society : Series B (Statistical Methodology) 68*(3), 411–436.

(Demiriz et al., 1999) A. Demiriz, K. P. Bennett, & M. J. Embrechts, 1999. Semi-supervised clustering using genetic algorithms. *Artificial neural networks in engineering (ANNIE-99)*, 809–814.

(Dempster et al., 1977) A. P. Dempster, N. M. Laird, D. B. Rubin, et al., 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society 39*(1), 1–38.

(Deng et al., 2013) L. Deng, X. He, & J. Gao, 2013. Deep stacking networks for information retrieval. Dans les actes de *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 3153–3157. IEEE.

(Deng et Yu, 2011) L. Deng & D. Yu, 2011. Deep convex net : A scalable architecture for speech pattern classification. Dans les actes de *Proceedings of the Interspeech*.

(Doddington et al., 2000) G. R. Doddington, M. A. Przybocki, A. F. Martin, & D. A. Reynolds, 2000. The nist speaker recognition evaluation–overview, methodology, systems, results, perspective. *Speech Communication 31*(2), 225–254.

(Dong et al., 2011) T. Dong, W. Shang, & H. Zhu, 2011. An improved algorithm of bayesian text categorization. *Journal of Software 6*(9), 1837–1843.

(Dong et Zhaohui, 2001) X. Dong & W. Zhaohui, 2001. Speaker recognition using continuous density support vector machines. *Electronics letters 37*(17), 1099–1101.

(Dumais, 1993) S. Dumais, 1993. Lsi meets trec : A status report. Dans les actes de *Proceedings of the first Text REtrieval Conference, TREC1*, 137–152.

(Dumais, 1994) S. Dumais, 1994. Latent semantic indexing (lsi) and trec-2. *NIST SPECIAL PUBLICATION SP*, 105–105.

(Dumais, 1991) S. T. Dumais, 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers 23*(2), 229–236.

(Ell, 1992) T. A. Ell, 1992. Hypercomplex spectral transformations.

(Ell et Sangwine, 2007) T. A. Ell & S. J. Sangwine, 2007. Hypercomplex fourier transforms of color images. *Image Processing, IEEE Transactions on 16*(1), 22–35.

(Erosheva et al., 2004) E. Erosheva, S. Fienberg, & J. Lafferty, 2004. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America 101*(Suppl 1), 5220–5227.

(Fan et al., 2006) W. Fan, L. Wallace, S. Rich, & Z. Zhang, 2006. Tapping the power of text mining. *Communications of the ACM 49*(9), 76–82.

(Fei-Fei et Perona, 2005) L. Fei-Fei & P. Perona, 2005. A bayesian hierarchical model for learning natural scene categories. Dans les actes de *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Volume 2, 524–531. IEEE.

(Feinerer, 2008) I. Feinerer, 2008. *A text mining framework in R and its applications*. Thèse de Doctorat, WU Vienna University of Economics and Business.

(Fergus et al., 2005) R. Fergus, L. Fei-Fei, P. Perona, & A. Zisserman, 2005. Learning object categories from google's image search. Dans les actes de *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, Volume 2, 1816–1823. IEEE.

(Foltz et Dumais, 1992) P. Foltz & S. Dumais, 1992. Personalized information delivery : An analysis of information filtering methods. *Communications of the ACM 35*(12), 51–60.

(Franco-Pedroso et al., 2010) J. Franco-Pedroso, I. Lopez-Moreno, D. T. Toledano, & J. Gonzalez-Rodriguez, 2010. Atvs-uam system description for the audio segmentation and speaker diarization albayzin 2010 evaluation. Dans les actes de *Iberian SLTech Workshop*, 415–418.

(Frobenius, 1877) G. Frobenius, 1877. Ueber lineare substitutionen und bilineare formen. *Journal für die reine und angewandte Mathematik 84*, 1–63.

(Froissar, 2007) P. Froissar, 2007. Buzz, bouffées d'audience et rumeur sur internet.

(Funda et al., 1990) J. Funda, R. H. Taylor, & R. P. Paul, 1990. On homogeneous transforms, quaternions, and computational efficiency. *Robotics and Automation, IEEE Transactions on 6*(3), 382–388.

(Furnas et al., 1988) G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, & K. E. Lochbaum, 1988. Information retrieval using a singular value decomposition model of latent semantic structure. Dans les actes de *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, 465–480. ACM.

(Gauvain et al., 1998) J.-L. Gauvain, L. Lamel, & G. Adda, 1998. Partitioning and transcription of broadcast news data. Dans les actes de *ICSLP*, Volume 98, 1335–1338.

(Gelman et al., 2013) A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, & D. B. Rubin, 2013. *Bayesian data analysis*. CRC press.

(Geman et Geman, 1984) S. Geman & D. Geman, 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (6), 721–741.

(Georgescul et al., 2006) M. Georgescul, A. Clark, & S. Armstrong, 2006. Word distributions for thematic segmentation in a support vector machine approach. Dans les actes de *Conference on Computational Natural Language Learning*, 101–108.

(Gildea et Hofmann, 1999) D. Gildea & T. Hofmann, 1999. Topic-based language models using em. *History 11111*(11111), 11111.

(Girolami et Kabán, 2003) M. Girolami & A. Kabán, 2003. On an equivalence between plsi and lda. Dans les actes de *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 433–434. ACM.

(Gish et al., 1991) H. Gish, M.-H. Siu, & R. Rohlicek, 1991. Segregation of speakers for speech recognition and speaker identification. Dans les actes de *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 873–876. IEEE.

(Glembek et al., 2009) O. Glembek, L. Burget, N. Dehak, N. Brummer, & P. Kenny, 2009. Comparison of scoring methods used in speaker recognition with joint factor analysis. Dans les actes de *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 4057–4060. IEEE.

(Goldenberg et al., 2001) J. Goldenberg, B. Libai, & E. Muller, 2001. Using complex systems analysis to advance marketing theory development : Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review 9*, 1–18.

(Golder et Huberman, 2007) S. Golder & B. Huberman, 2007. The structure of collaborative tagging systems.

(Golub et Van Loan, 1989) G. H. Golub & C. F. Van Loan, 1989. *Matrix computations*, Volume 3. JHU Press.

(Griffiths et Steyvers, 2002) T. Griffiths & M. Steyvers, 2002. A probabilistic approach to semantic representation. Dans les actes de *24th annual conference of the cognitive science society*, 381–386. Citeseer.

(Griffiths et Steyvers, 2004) T. L. Griffiths & M. Steyvers, 2004. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America 101*(Suppl 1), 5228–5235.

(Griffiths et al., 2004) T. L. Griffiths, M. Steyvers, D. M. Blei, & J. B. Tenenbaum, 2004. Integrating topics and syntax. Dans les actes de *NIPS*, 537–544.

(Gruber et al., 2007) A. Gruber, Y. Weiss, & M. Rosen-Zvi, 2007. Hidden topic markov models. Dans les actes de *International Conference on Artificial Intelligence and Statistics*, 163–170.

(Gunal, 2012) S. Gunal, 2012. Hybrid feature selection for text classification. *Turkish Journal of Electrical Engineering & Computer Sciences 20*(2), 1296–1311.

(Guo et Zhu, 2011) L.-Q. Guo & M. Zhu, 2011. Quaternion fourier–mellin moments for color images. *Pattern Recognition 44*(2), 187–195.

(Hamilton, 1866) S. Hamilton, 1866. *Elements of quaternions*. Longmans, Green, & co.

(Harish et al., 2010) B. Harish, D. Guru, & S. Manjunath, 2010. Representation and classification of text documents : A brief review. *International Journal of Computer Applications IJCA* (2), 110–119.

(Hazen, 2011) T. Hazen, 2011. Topic identification. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, 319–356.

(Hearst, 1999) M. A. Hearst, 1999. Untangling text data mining. Dans les actes de *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 3–10. Association for Computational Linguistics.

(Heck et al., 2000) L. P. Heck, Y. Konig, M. K. Sönmez, & M. Weintraub, 2000. Robustness to telephone handset distortion in speaker recognition by discriminative feature design. *Speech Communication 31*(2), 181–192.

(Heidel et al., 2007) A. Heidel, H.-a. Chang, & L.-s. Lee, 2007. Language model adaptation using latent dirichlet allocation and an efficient topic inference algorithm. Dans les actes de *INTERSPEECH*, 2361–2364.

(Heinrich, 2005) G. Heinrich, 2005. Parameter estimation for text analysis. Rapport technique, Technical report.

(Hinton, 1986) G. E. Hinton, 1986. Learning distributed representations of concepts. Dans les actes de *Proceedings of the eighth annual conference of the cognitive science society*, Volume 1, 12. Amherst, MA.

(Hitzer, 2013) E. Hitzer, 2013. Quaternionic fourier-mellin transform. *arXiv preprint arXiv :1306.1669*.

(Hitzer et Sangwine, 2013) E. Hitzer & S. J. Sangwine, 2013. *Quaternion and Clifford Fourier transforms and wavelets*. Springer.

(Hoffman et al., 2010) M. D. Hoffman, D. M. Blei, & F. R. Bach, 2010. Online learning for latent dirichlet allocation. Dans les actes de *NIPS*, Volume 2, 5.

(Hofmann, 2001) D. C. T. Hofmann, 2001. The missing link-a probabilistic model of document content and hypertext connectivity. Dans les actes de *Proceedings of the 2000 Conference on Advances in Neural Information Processing Systems. The MIT Press*, 430–436.

(Hofmann, 1999a) T. Hofmann, 1999a. Probabilistic latent semantic analysis. Dans les actes de *Proc. of Uncertainty in Artificial Intelligence, UAI ' 99*, 21. Citeseer.

(Hofmann, 1999b) T. Hofmann, 1999b. Probabilistic latent semantic indexing. Dans les actes de *22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50–57. ACM.

(Hofmann et al., 1999) T. Hofmann, J. Puzicha, & M. I. Jordan, 1999. Learning from dyadic data. *Advances in neural information processing systems*, 466–472.

(Hotelling, 1933) H. Hotelling, 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology 24*(6), 417.

(hua Yeh et hsing Chen, 2010) J. hua Yeh & C. hsing Chen, 2010. Protein remote homology detection based on latent topic vector model. Dans les actes de *International Conference on Networking and Information Technology (ICNIT)*, 456–460.

(infoGuerre, 2000) infoGuerre, 2000. La guerre de l ?information du faible au fort. http://strategique.free.fr/archives/textes/infog/archives_infog_06.htm.

(Iwata et al., 2007) T. Iwata, K. Saito, N. Ueda, S. Stromsten, T. L. Griffiths, & J. B. Tenenbaum, 2007. Parametric embedding for class visualization. *Neural Computation 19*(9), 2536–2556.

(Joachims, 1999) T. Joachims, 1999. Transductive inference for text classification using support vector machines. Dans les actes de *Machine learning-international workshop then conference*, 200–209. Morgan Kaufmann Publishers, Inc.

(Jolliffe, 1986) I. T. Jolliffe, 1986. Principal component analysis. New York : Springer-Verlag.

(Jones, 1972) K. S. Jones, 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation 28*(1), 11–21.

(Jordan et al., 2004) M. I. Jordan et al., 2004. Graphical models. *Statistical Science 19*(1), 140–155.

(Kanerva et al., 2000) P. Kanerva, J. Kristofersson, & A. Holst, 2000. Random indexing of text samples for latent semantic analysis. Dans les actes de *Proceedings of the 22nd annual conference of the cognitive science society*, Volume 1036. Citeseer.

(Kantor et al., 1989) I. Kantor, A. Solodovnikov, & A. Shenitzer, 1989. *Hypercomplex numbers : an elementary introduction to algebras*. Springer-Verlag.

(Kempe et al., 2003) D. Kempe, J. Kleinberg, & é. Tardos, 2003. Maximizing the spread of influence through a social network. Dans les actes de *ACM SIGKDD International Conf. on Knowledge discovery and data mining*, 137–146.

(Kenny, 2008) P. Kenny, 2008. Bayesian analysis of speaker diarization with eigenvoice priors. *CRIM, Montreal, Technical Report*.

(Kenny, 2010) P. Kenny, 2010. Bayesian speaker verification with heavy-tailed priors. Dans les actes de *Odyssey*, 14.

(Kenny et al., 2005) P. Kenny, G. Boulianne, P. Ouellet, & P. Dumouchel, 2005. Factor Analysis Simplified. Dans les actes de *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005)*, Volume 1, 637–640.

(Kenny et al., 2007) P. Kenny, G. Boulianne, P. Ouellet, & P. Dumouchel, 2007. Joint factor analysis versus eigenchannels in speaker recognition. *Audio, Speech, and Language Processing, IEEE Transactions on 15*(4), 1435–1447.

(Kenny et Dumouchel, 2004) P. Kenny & P. Dumouchel, 2004. Experiments in speaker verification using factor analysis likelihood ratios. Dans les actes de *ODYSSEY04-The Speaker and Language Recognition Workshop*.

(Kenny et al., 2008) P. Kenny, P. Ouellet, N. Dehak, V. Gupta, & P. Dumouchel, 2008. A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing 16*(5), 980–988.

(Kim et al., 2009) S. Kim, S. Narayanan, & S. Sundaram, 2009. Acoustic topic model for audio information retrieval. Dans les actes de *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 37–40.

(Kim et al., 2003) Y.-S. Kim, J.-H. Chang, & B.-T. Zhang, 2003. An empirical study on dimensionality optimization in text mining for linguistic knowledge acquisition. Dans les actes de *Advances in Knowledge Discovery and Data Mining*, 111–116. Springer.

(Kiritchenko et Matwin, 2001) S. Kiritchenko & S. Matwin, 2001. Email classification with co-training. Dans les actes de *Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative research*, 8. IBM Press.

(Koço et al., 2012) S. Koço, C. Capponi, & F. Béchet, 2012. Applying multiview learning algorithms to human-human conversation classification. Dans les actes de *Portland, September 2012 (INTERSPEECH) ISCA*.

(Kruschwitz, 2005) U. Kruschwitz, 2005. *Intelligent document retrieval : exploiting markup structure*, Volume 17. Springer.

(Kuipers, 1999) J. B. Kuipers, 1999. *Quaternions and rotation sequences*. Princeton university press Princeton, NJ, USA :.

(Kullback, 1987) S. Kullback, 1987. The kullback-leibler distance.

(Lan et al., 2005) M. Lan, C.-L. Tan, H.-B. Low, & S.-Y. Sung, 2005. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. Dans les actes de *International Conference on World Wide Web*, 1032–1033.

(Landauer et Dumais, 1997) T. Landauer & S. Dumais, 1997. A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review 104*(2), 211.

(Landauer et al., 1997) T. Landauer, D. Laham, B. Rehder, & M. Schreiner, 1997. How well can passage meaning be derived without using word order ? a comparison of latent semantic analysis and humans. Dans les actes de *Proceedings of the nineteenth annual conference of the Cognitive Science Society : August 7-10, 1997, Stanford University, Stanford, CA*, Volume 10, 412. Lawrence Erlbaum.

(Larson et al., 2010) M. Larson, E. Newman, & G. J. Jones, 2010. Overview of videoclef 2009 : New perspectives on speech-based multimedia content enrichment. Dans les actes de *Multilingual Information Access Evaluation II. Multimedia Experiments*, 354–368. Springer.

(Le Bihan et al., 2006) N. Le Bihan, S. Buchholz, B. ENSIEG, C. Kiel, & N. Le-Bihan, 2006. Quaternionic independent component analysis using hypercomplex nonlinearities. Dans les actes de *Proc. 7th IMA Conf. Math. Signal Process*, 1–3.

(Le Bihan et Sangwine, 2003a) N. Le Bihan & S. J. Sangwine, 2003a. Color image decomposition using quaternion singular value decomposition. Dans les actes de *Visual Information Engineering, 2003. VIE 2003. International Conference on*, 113–116. IET.

(Le Bihan et Sangwine, 2003b) N. Le Bihan & S. J. Sangwine, 2003b. Quaternion principal component analysis of color images. Dans les actes de *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, Volume 1, I–809. IEEE.

(Lecouteux et al., 2009) B. Lecouteux, G. Linares, & B. Favre, 2009. Combined low level and high level features for out-of-vocabulary word detection. Dans les actes de *INTERSPEECH*, 1187–1190.

(Li et al., 2012) P. Li, Y. Fu, U. Mohammed, J. H. Elder, & S. J. Prince, 2012. Probabilistic models for inference about identity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 34*(1), 144–157.

(Li et McCallum, 2006) W. Li & A. McCallum, 2006. Pachinko allocation : Dag-structured mixture models of topic correlations.

(Li, 2013) Y. N. Li, 2013. Quaternion polar harmonic transforms for color images. *Signal Processing Letters, IEEE 20*(8), 803–806.

(Lienou et al., 2010) M. Lienou, H. Maitre, & M. Datcu, 2010. Semantic annotation of satellite images using latent dirichlet allocation. *Geoscience and Remote Sensing Letters, IEEE 7*(1), 28–32.

(Linarès et al., 2007) G. Linarès, P. Nocéra, D. Massonie, & D. Matrouf, 2007. The lia speech recognition system : from 10xrt to 1xrt. Dans les actes de *Text, Speech and Dialogue*, 302–308. Springer.

(Liu et al., 2008) F. Liu, F. Liu, & Y. Liu, 2008. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. Dans les actes de *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, 181–184. IEEE.

(Liu et al., 2011) X. Liu, B. Huet, & R. Troncy, 2011. Eurecom@ mediaeval 2011 social event detection task. Dans les actes de *MediaEval*.

(Lu et Zhai, 2008) Y. Lu & C. Zhai, 2008. Opinion integration through semi-supervised topic modeling. Dans les actes de *Proceedings of the 17th international conference on World Wide Web*, 121–130. ACM.

(MacKay, 2003) D. J. MacKay, 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

(Magrin-Chagnolleau et al., 1999) I. Magrin-Chagnolleau, A. E. Rosenberg, & S. Parthasarathy, 1999. Detection of target speakers in audio databases. Dans les actes de *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, Volume 2, 821–824. IEEE.

(Mami et Charlet, 2002) Y. Mami & D. Charlet, 2002. Speaker identification by location in an optimal space of anchor models. Dans les actes de *INTERSPEECH*.

(Manning et al., 2008) C. D. Manning, P. Raghavan, & H. Schütze, 2008. *Introduction to information retrieval*, Volume 1. Cambridge university press Cambridge.

(Manning et Schütze, 1999) C. D. Manning & H. Schütze, 1999. *Foundations of statistical natural language processing*. MIT press.

(Martınez et al., 2011) D. Martınez, O. Plchot, L. Burget, O. Glembek, & P. Matejka, 2011. Language recognition in ivectors space. *INTERSPEECH*, 861–864.

(Matrouf et al., 2007) D. Matrouf, N. Scheffer, B. G. Fauve, & J.-F. Bonastre, 2007. A straightforward and efficient implementation of the factor analysis model for speaker verification. Dans les actes de *Interspeech'07*, 1242–1245.

(Matrouf et al., 2011) D. Matrouf, F. Verdet, M. Rouvier, J.-F. Bonastre, & G. Linarès, 2011. Modeling nuisance variabilities with factor analysis for gmm-based audio pattern classification. *Computer Speech & Language 25*(3), 481–498.

(Maza et al., 2011) B. Maza, M. El-Beze, G. Linares, & R. Mori, 2011. On the use of linguistic features in an automatic system for speech analytics of telephone conversations. Dans les actes de *Interspeech'11*.

(McCallum, 2002) A. McCallum, 2002. Mallet : A machine learning for language toolkit.

(McLachlan et Basford, 1988) G. J. McLachlan & K. E. Basford, 1988. Mixture models. inference and applications to clustering. *Statistics : Textbooks and Monographs, New York : Dekker, 1988 1*.

(Mei et al., 2007) Q. Mei, X. Ling, M. Wondra, H. Su, & C. Zhai, 2007. Topic sentiment mixture : modeling facets and opinions in weblogs. Dans les actes de *Proceedings of the 16th international conference on World Wide Web*, 171–180. ACM.

(Melamed et Gilbert, 2011) I. Melamed & M. Gilbert, 2011. Speech analytics. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, 397–416.

(Mennesson et al., 2011) J. Mennesson, C. Saint-Jean, & L. Mascarilla, 2011. Color object recognition based on a clifford fourier transform. Dans les actes de *Guide to geometric algebra in practice*, 175–191. Springer.

(Mennesson et al., 2014) J. Mennesson, C. Saint-Jean, & L. Mascarilla, 2014. Color fourier–mellin descriptors for image recognition. *Pattern Recognition Letters 40*, 27–35.

(Mikolov et al., 2013) T. Mikolov, K. Chen, G. Corrado, & J. Dean, 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

(Miller, 1995) G. A. Miller, 1995. Wordnet : a lexical database for english. *Communications of the ACM 38*(11), 39–41.

(Minka et Lafferty, 2002) T. Minka & J. Lafferty, 2002. Expectation-propagation for the generative aspect model. Dans les actes de *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, 352–359. Morgan Kaufmann Publishers Inc.

(Mirsky, 1960) L. Mirsky, 1960. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics 11*(1), 50–59.

(Morchid et al., 2014b) M. Morchid, R. Dufour, P.-M. Bousquet, M. Bouallegue, G. Linarès, & R. De Mori, 2014b. Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule. Dans les actes de *International Conference on Acoustic, Speech and Signal Processing (ICASSP) 2014*. IEEE.

(Morchid et al., 2013a) M. Morchid, R. Dufour, & G. Linarès, 2013a. Event detection from image hosting services by slightly-supervised multi-span context models. Dans les actes de *International Workshop on Content-Based Multimedia Indexing (CBMI) 2013*. IEEE.

(Morchid et al., 2013b) M. Morchid, R. Dufour, & G. Linarès, 2013b. Thematic representation of short text messages with latent topics : Application in the twitter context. Dans les actes de *Conference of the Pacific Association for Computational Linguistics (PACLING) 2013*.

(Morchid et al., 2014a) M. Morchid, R. Dufour, & G. Linarès, 2014a. A lda-based topic classification approach from highly imperfect automatic transcriptions. Dans les actes de *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC) 2014*.

(Morchid et Linarès, 2013a) M. Morchid & G. Linarès, 2013a. Lda-based method for automatic tagging of youtube videos. Dans les actes de *nternational Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS) 2013*. IEEE.

(Morchid et Linarès, 2013b) M. Morchid & G. Linarès, 2013b. Prédiction des buzz sur twitter. Dans les actes de *COnférence en Recherche d'Information et Applications (CORIA) 2013*.

(Morchid et al., 2014c) M. Morchid, G. Linarès, & R. Dufour, 2014c. Characterizing and predicting bursty events : The buzz case study on twitter. Dans les actes de *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC) 2014*.

(Mrva et Woodland, 2004) D. Mrva & P. C. Woodland, 2004. A plsa-based language model for conversational telephone speech. Dans les actes de *INTERSPEECH*.

(Müller et al., 1997) K. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, & V. Vapnik, 1997. Predicting time series with support vector machines. *ICANN'97*, 999–1004.

(Murray et Salakhutdinov, 2008) I. Murray & R. Salakhutdinov, 2008. Evaluating probabilities under high-dimensional latent variable models. Dans les actes de *NIPS*, 1137–1144.

(Nallapati et Cohen, 2008) R. Nallapati & W. W. Cohen, 2008. Link-plsa-lda : A new unsupervised model for topics and influence of blogs. Dans les actes de *ICWSM*.

(Nathan, 2009) S. Nathan, 2009. Improving sentimental classifications using contextual sentences lexical base (pp. 1-8).

(Neapolitan et al., 2003) R. E. Neapolitan et al., 2003. *Learning bayesian networks*, Volume 1. Prentice Hall Upper Saddle River.

(Newton et Raftery, 1994) M. A. Newton & A. E. Raftery, 1994. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–48.

(Niebles et al., 2008) J. C. Niebles, H. Wang, & L. Fei-Fei, 2008. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision* *79*(3), 299–318.

(Niquet, 2000) V. Niquet, 2000. La guerre de l'information dans la pensée stratégique chinoise. http://www.infoguerre.fr/doctrines/la-guerre-de-l-information-dans-la-pensee-strategique-chinoise/.

(NIST, 2005) NIST, 2005. The nist year 2005 speaker recognition evaluation plan. http://www.itl.nist.gov/iad/mig//tests/spk/2005/sre-05_evalplan-v6.pdf. [Online ; accessed 2005].

(Nitta, 1995) T. Nitta, 1995. A quaternary version of the back-propagation algorithm. Dans les actes de *Proceedings of the IEEE International Conference on Neural Networks*, Volume 5, 2753–2756.

(Niu et Shi, 2010) L. Niu & Y. Shi, 2010. Semi-supervised plsa for document clustering. Dans les actes de *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, 1196–1203. IEEE.

(Papadimitriou et al., 1998) C. H. Papadimitriou, H. Tamaki, P. Raghavan, & S. Vempala, 1998. Latent semantic indexing : A probabilistic analysis. Dans les actes de *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, 159–168. ACM.

(Papadopoulos et al., 2011a) S. Papadopoulos, R. Troncy, V. Mezaris, B. Huet, & I. Kompatsiaris, 2011a. Social Event Detection at MediaEval 2011 : Challenges, Dataset and Evaluation. Dans les actes de *MediaEval 2011 Workshop*, Pisa, Italy.

(Papadopoulos et al., 2011b) S. Papadopoulos, R. Troncy, V. Mezaris, B. Huet, & I. Kompatsiaris, 2011b. Social event detection at mediaeval 2011 : Challenges, dataset and evaluation. Dans les actes de *MediaEval*.

(Pearson et Bisset, 1994) J. Pearson & D. Bisset, 1994. Neural networks in the clifford domain. Dans les actes de *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*, Volume 3, 1465–1469. IEEE.

(Pearson, 1901) K. Pearson, 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* *2*(11), 559–572.

(Pei et al., 2003) S.-C. Pei, J.-H. Chang, & J.-J. Ding, 2003. Quaternion matrix singular value decomposition and its applications for color image processing. Dans les actes de *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, Volume 1, I–805. IEEE.

(Pei et Cheng, 1997) S.-C. Pei & C.-M. Cheng, 1997. A novel block truncation coding of color images using a quaternion-moment-preserving principle. *Communications, IEEE Transactions on* *45*(5), 583–595.

(Pei et Cheng, 1999) S.-C. Pei & C.-M. Cheng, 1999. Color image processing by using binary quaternion-moment-preserving thresholding technique. *Image Processing, IEEE Transactions on* *8*(5), 614–628.

(Petridis et Perantonis, 2004) S. Petridis & S. J. Perantonis, 2004. On the relation between discriminant analysis and mutual information for supervised linear feature extraction. *Pattern Recognition* *37*(5), 857–874.

(Phan et al., 2008) X.-H. Phan, L.-M. Nguyen, & S. Horiguchi, 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Dans les actes de *Proceedings of the 17th international conference on World Wide Web*, 91–100. ACM.

(Piwowarski, 2013) B. Piwowarski, 2013. Méthodologie pour une représentation multi-dimensionnelle des documents.

(Povey et al., 2010) D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, et al., 2010. Subspace gaussian mixture models for speech recognition. Dans les actes de *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 4330–4333. IEEE.

(Purver, 2011) M. Purver, 2011. Topic segmentation. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, 291–317.

(Ramage et al., 2009) D. Ramage, D. Hall, R. Nallapati, & C. D. Manning, 2009. Labeled lda : A supervised topic model for credit attribution in multi-labeled corpora. Dans les actes de *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1-Volume 1*, 248–256. Association for Computational Linguistics.

(Rattenbury et al., 2007) T. Rattenbury, N. Good, & M. Naaman, 2007. Towards automatic extraction of event and place semantics from flickr tags. Dans les actes de *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 103–110. ACM.

(Reynolds et al., 2000) D. Reynolds, T. Quatieri, & R. Dunn, 2000. Speaker verification using adapted gaussian mixture models. *Digital signal processing 10*(1-3), 19–41.

(Reynolds et Rose, 1995) D. Reynolds & R. Rose, 1995. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on 3*(1), 72–83.

(Reynolds et al., 1998) D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. McLaughlin, & M. A. Zissman, 1998. Blind clustering of speech utterances based on speaker and language characteristics. Dans les actes de *ICSLP*.

(Rizo-Rodríguez et al., 2013) D. Rizo-Rodríguez, H. Méndez-Vázquez, & E. García-Reyes, 2013. Illumination invariant face recognition using quaternion-based correlation filters. *Journal of mathematical imaging and vision 45*(2), 164–175.

(Robertson, 2004) S. Robertson, 2004. Understanding inverse document frequency : on theoretical arguments for idf. *Journal of Documentation 60*(5), 503–520.

(Robertson et al., 2000) S. E. Robertson, S. Walker, & M. Beaulieu, 2000. Experimentation as a way of life : Okapi at trec. *Information Processing & Management 36*(1), 95–108.

(Romero et al., 2011) D. Romero, B. Meeder, & J. Kleinberg, 2011. Differences in the mechanics of information diffusion across topics : idioms, political hashtags, and complex contagion on twitter. Dans les actes de *ACM International Conf. on World Wide Web*, 695–704.

(Rose et al., 1990) K. Rose, E. Gurewitz, & G. Fox, 1990. A deterministic annealing approach to clustering. *Pattern Recognition Letters 11*(9), 589–594.

(Rosen-Zvi et al., 2004) M. Rosen-Zvi, T. Griffiths, M. Steyvers, & P. Smyth, 2004. The author-topic model for authors and documents. Dans les actes de *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 487–494. AUAI Press.

(Rosenberg et al., 1998) A. Rosenberg, O. Siohan, & S. Parathasarathy, 1998. Speaker verification using minimum verification error training. Dans les actes de *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, Volume 1, 105–108. IEEE.

(Rouvier et al., 2011) M. Rouvier, M. Bouallegue, D. Matrouf, & G. Linarès, 2011. Factor analysis based session variability compensation for automatic speech recognition. Dans les actes de *ASRU*, 141–145.

(Rouvier et al., 2009) M. Rouvier, G. Linares, & D. Matrouf, 2009. Robust audio-based classification of video genre. Dans les actes de *Interspeech'09*, 1159–1162.

(Russell et al., 2006) B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, & A. Zisserman, 2006. Using multiple segmentations to discover objects and their extent in image collections. Dans les actes de *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, Volume 2, 1605–1614. IEEE.

(Sahlgren et Cöster, 2004) M. Sahlgren & R. Cöster, 2004. Using bag-of-concepts to improve the performance of support vector machines in text categorization. Dans les actes de *Proceedings of the 20th international conference on Computational Linguistics*, 487. Association for Computational Linguistics.

(Salton, 1971) G. Salton, 1971. The smart retrieval system ?experiments in automatic document processing.

(Salton, 1989a) G. Salton, 1989a. Automatic text processing-the analysis, transformation and retrieval of information by computer.

(Salton, 1989b) G. Salton, 1989b. Automatic text processing : the transformation. *Analysis and Retrieval of Information by Computer*.

(Salton et Buckley, 1988) G. Salton & C. Buckley, 1988. Term-weighting approaches in automatic text retrieval* 1. *Information processing & management* *24*(5), 513–523.

(Salton et McGill, 1983) G. Salton & M. J. McGill, 1983. Introduction to modern information retrieval.

(Salton et Yang, 1973) G. Salton & C.-S. Yang, 1973. On the specification of term values in automatic indexing. *Journal of documentation* *29*(4), 351–372.

(Salton et al., 1975) G. Salton, C.-S. Yang, & C. T. Yu, 1975. A theory of term importance in automatic text analysis. *Journal of the American society for Information Science* *26*(1), 33–44.

(Sangwine, 1996) S. J. Sangwine, 1996. Fourier transforms of colour images using quaternion or hypercomplex, numbers. *Electronics letters* *32*(21), 1979–1980.

(SanJuan et al., 2012) E. SanJuan, V. Moriceau, X. Tannier, P. Bellot, & J. Mothe, 2012. Overview of the inex 2012 tweet contextualization track. Dans les actes de *INEX 2012 conference book*, 148.

(Sarikaya et al., 2011) R. Sarikaya, G. E. Hinton, & B. Ramabhadran, 2011. Deep belief nets for natural language call-routing. Dans les actes de *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 5680–5683. IEEE.

(Savoy et Dolamic, 2008) J. Savoy & L. Dolamic, 2008. Variations autour de tf idf et du moteur lucene. *9es Journées internationales d ?Analyse statistique des Données Textuelles*, 1047–1058.

(Schapire et Singer, 2000) R. E. Schapire & Y. Singer, 2000. Boostexter : A boosting-based system for text categorization. *Machine learning 39*(2-3), 135–168.

(Schmidt et Gish, 1996) M. Schmidt & H. Gish, 1996. Speaker identification via support vector classifiers. Dans les actes de *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, Volume 1, 105–108. IEEE.

(Sciavicco et Villani, 2009) L. Sciavicco & L. Villani, 2009. *Robotics : modelling, planning and control*. Springer.

(Sheeba et Vivekanandan, 2012) J. I. Sheeba & K. Vivekanandan, 2012. Article : Improved keyword and keyphrase extraction from meeting transcripts. *International Journal of Computer Applications 52*(13), 11–15.

(Shuster, 1993) M. D. Shuster, 1993. A survey of attitude representations. *Navigation 8*(9).

(Siegler et al., 1997) M. A. Siegler, U. Jain, B. Raj, & R. M. Stern, 1997. Automatic segmentation, classification and clustering of broadcast news audio. Dans les actes de *Proc. DARPA Broadcast News Workshop*, 11.

(Sivic et al., 2005) J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, & W. T. Freeman, 2005. Discovering objects and their location in images. Dans les actes de *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, Volume 1, 370–377. IEEE.

(Snyder et Barzilay, 2007) B. Snyder & R. Barzilay, 2007. Multiple aspect ranking using the good grief algorithm. Dans les actes de *HLT-NAACL*, 300–307.

(Stein et Schmid, 1995) A. Stein & H. Schmid, 1995. Etiquetage morphologique de textes français avec un arbre de décisions. *Traitement automatique des langues 36*(1-2), 23–35.

(Steyvers et al., 2004) M. Steyvers, P. Smyth, M. Rosen-Zvi, & T. Griffiths, 2004. Probabilistic author-topic models for information discovery. Dans les actes de *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 306–315. ACM.

(Stock, 2007) W. G. Stock, 2007. *Information retrieval : Informationen suchen und finden ;[Lehrbuch]*, Volume 1. Oldenbourg Verlag.

(Story, 1996) R. Story, 1996. An explanation of the effectiveness of latent semantic indexing by means of a bayesian regression model. *Information Processing & Management 32*(3), 329–344.

(Subakan et Vemuri, 2011) Ö. N. Subakan & B. C. Vemuri, 2011. A quaternion framework for color image smoothing and segmentation. *International Journal of Computer Vision 91*(3), 233–250.

(Sun et al., 2011) Y. Sun, S. Chen, & B. Yin, 2011. Color face recognition based on quaternion matrix representation. *Pattern Recognition Letters 32*(4), 597–605.

(Susbielle, 2006) J.-F. Susbielle, 2006. *Chine-USA, la guerre programmée*. First.

(Tang et al., 2009) S. Tang, J. Li, Y. Zhang, C. Xie, M. Li, Y. Liu, X. Hua, Y.-T. Zheng, J. Tang, & T.-S. Chua, 2009. Pornprobe : an lda-svm based pornography detection system. Dans les actes de *International Conference on Multimedia*, 1003–1004.

(Teh et al., 2004) Y. W. Teh, M. I. Jordan, M. J. Beal, & D. M. Blei, 2004. Sharing clusters among related groups : Hierarchical dirichlet processes. Dans les actes de *NIPS*.

(Tipping et Bishop, 1999) M. E. Tipping & C. M. Bishop, 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B 61*, 611–622.

(Titov et McDonald, 2008) I. Titov & R. McDonald, 2008. Modeling online reviews with multi-grain topic models. Dans les actes de *Proceedings of the 17th international conference on World Wide Web*, 111–120. ACM.

(Troncy et al., 2010) R. Troncy, B. Malocha, & A. T. Fialho, 2010. Linking events with media. Dans les actes de *Proceedings of the 6th International Conference on Semantic Systems*, 42. ACM.

(Tur et De Mori, 2011) G. Tur & R. De Mori, 2011. *Spoken language understanding : Systems for extracting semantic information from speech*. John Wiley & Sons.

(Tur et al., 2012) G. Tur, L. Deng, D. Hakkani-Tur, & X. He, 2012. Towards deeper understanding : deep convex networks for semantic utterance classification. Dans les actes de *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 5045–5048. IEEE.

(Van Asch, 2013) V. Van Asch, 2013. Macro-and micro-averaged evaluation measures [[basic draft]].

(Van Rijsbergen, 2004) C. J. Van Rijsbergen, 2004. *The geometry of information retrieval*, Volume 157. Cambridge University Press Cambridge.

(Vapnik, 1963) V. Vapnik, 1963. Pattern recognition using generalized portrait method. *Automation and Remote Control 24*, 774–780.

(Vesnicer et al., 2012) B. Vesnicer, J. Ž. Gros, N. Pavešić, & V. Štruc, 2012. Face recognition using simplified probabilistic linear discriminant analysis. *Int J Adv Robotic Sy 9*(180).

(Vía et al., 2011) J. Vía, D. P. Palomar, L. Vielva, & I. Santamaría, 2011. Quaternion ica from second-order statistics. *Signal Processing, IEEE Transactions on 59*(4), 1586–1600.

(Vogt et al., 2005) R. J. Vogt, B. J. Baker, & S. Sridharan, 2005. Modelling session variability in text independent speaker verification.

(Wallach, 2006) H. M. Wallach, 2006. Topic modeling : beyond bag-of-words. Dans les actes de *Proceedings of the 23rd international conference on Machine learning*, 977–984. ACM.

(Wallach et al., 2009) H. M. Wallach, I. Murray, R. Salakhutdinov, & D. Mimno, 2009. Evaluation methods for topic models. Dans les actes de *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105–1112. ACM.

(Wan et Renals, 2003) V. Wan & S. Renals, 2003. Svmsvm : support vector machine speaker verification methodology. Dans les actes de *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, Volume 2, II–221. IEEE.

(Wan et Renals, 2005) V. Wan & S. Renals, 2005. Speaker verification using sequence discriminant support vector machines. *Speech and Audio Processing, IEEE Transactions on 13*(2), 203–210.

(Wang et al., 2007) X. Wang, X. Ma, & E. Grimson, 2007. Unsupervised activity perception by hierarchical bayesian models. Dans les actes de *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 1–8. IEEE.

(Wang et McCallum, 2005) X. Wang & A. McCallum, 2005. A note on topical n-grams.

(Wang et McCallum, 2006) X. Wang & A. McCallum, 2006. Topics over time : a non-markov continuous-time model of topical trends. Dans les actes de *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 424–433. ACM.

(Ward, 1997) J. Ward, 1997. *Quaternions and Cayley numbers : Algebra and applications*, Volume 403. Springer.

(Wei et Croft, 2006) X. Wei & W. B. Croft, 2006. Lda-based document models for ad-hoc retrieval. Dans les actes de *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 178–185. ACM.

(Wei et Croft, 2007) X. Wei & W. B. Croft, 2007. Investigating retrieval performance with manually-built topic models. Dans les actes de *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, 333–349. Le Centre De Hautes Etudes Internationales D'Informatique Documentaire.

(Weiss et al., 2010) S. M. Weiss, N. Indurkhya, T. Zhang, & F. Damerau, 2010. *Text mining : predictive methods for analyzing unstructured information*. Springer.

(Wilkinson et al., 1971) J. H. Wilkinson, C. Reinsch, F. Bauer, A. Householder, F. Olver, H. Rutishauser, K. Samelson, & E. Stiefel, 1971. *Handbook for Automatic Computation : Volume II : Linear Algebra*. Springer Berlin Heidelberg.

(William et al., 1995) C. William et al., 1995. Fast effective rule induction. Dans les actes de *Twelfth International Conference on Machine Learning*, 115–123.

(Wong et al., 2007) K.-Y. K. Wong, T.-K. Kim, & R. Cipolla, 2007. Learning motion categories using both semantic and structural information. Dans les actes de *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 1–6. IEEE.

(Woodland et al., 1997) P. Woodland, M. Gales, D. Pye, & S. Young, 1997. The development of the 1996 htk broadcast news transcription system. Dans les actes de *DARPA speech recognition workshop*, 73–78. Morgan Kaufmann Pub.

(Wu et al., 2005) Y. Wu, X. Hu, D. Hu, T. Li, & J. Lian, 2005. Strapdown inertial navigation system algorithms based on dual quaternions. *Aerospace and Electronic Systems, IEEE Transactions on 41*(1), 110–132.

(Xi et al., 2013) Y. T. Xi, M. Paulik, V. R. Gadde, & A. Sankar, 2013. Kpcatcher–a keyphrase extraction system for enterprise videos.

(Xian et al., 2004) B. Xian, M. S. de Queiroz, D. Dawson, & I. Walker, 2004. Task-space tracking control of robot manipulators via quaternion feedback. *Robotics and Automation, IEEE Transactions on 20*(1), 160–167.

(Xing et al., 2002) E. P. Xing, M. I. Jordan, S. Russell, & A. Ng, 2002. Distance metric learning with application to clustering with side-information. Dans les actes de *Advances in neural information processing systems*, 505–512.

(Xue et al., 2008) G.-R. Xue, W. Dai, Q. Yang, & Y. Yu, 2008. Topic-bridged plsa for cross-domain text classification. Dans les actes de *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 627–634. ACM.

(Yang et Counts, 2010) J. Yang & S. Counts, 2010. Predicting the speed, scale, and range of information diffusion in twitter. *Proc. ICWSM*.

(Yang et al., 1998) Y. Yang, T. Pierce, & J. Carbonell, 1998. A study of retrospective and on-line event detection. Dans les actes de *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 28–36. ACM.

(Yuan et al., 2012) G.-X. Yuan, C.-H. Ho, & C.-J. Lin, 2012. Recent advances of large-scale linear classification. *100*(9), 2584–2603.

(Yuan, 1988) J. Yuan, 1988. Closed-loop manipulator control using quaternion feedback. *Robotics and Automation, IEEE Journal of 4*(4), 434–440.

(Zavitsanos et al., 2008) E. Zavitsanos, S. Petridis, G. Paliouras, & G. A. Vouros, 2008. Determining automatically the size of learned ontologies. Dans les actes de *ECAI*, Volume 178, 775–776.

(Zhai et Lafferty, 2001) C. Zhai & J. Lafferty, 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. Dans les actes de *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 334–342. ACM.

(Zhang, 1997) F. Zhang, 1997. Quaternions and matrices of quaternions. *Linear algebra and its applications 251*, 21–57.

(Zhang et Gong, 2010) J. Zhang & S. Gong, 2010. Action categorization by structural probabilistic latent semantic analysis. *Computer Vision and Image Understanding 114*(8), 857–864.

(Zhao et Dong, 2012) X. Zhao & Y. Dong, 2012. Variational bayesian joint factor analysis models for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on 20*(3), 1032–1042.

(Zhu et Lin, 2013) W. Zhu & Y. Lin, 2013. Using gini-index for feature weighting in text categorization. *Journal of Computational Information Systems 9*(14), 5819–5826.

(Zhuang et al., 2009) L. Zhuang, L. She, Y. Jiang, K. Tang, & N. Yu, 2009. Image classification via semi-supervised plsa. Dans les actes de *Image and Graphics, 2009. ICIG'09. Fifth International Conference on*, 205–208. IEEE.

(Zrigui et al., 2012) M. Zrigui, R. Ayadi, M. Mars, & M. Maraoui, 2012. Arabic text classification framework based on latent dirichlet allocation. *CIT 20*(2), 125–140.

Glossaire et Acronymes

ACP	Analyse en Composantes Principales ou Principal Component Analysis. 50, 51, 62–66, 106, 108
BVB	Batch Variational Bayesian. 43
DCLM	Dirichlet class language model. 44
EM	L'algorithme espérance-maximisation (en anglais Expectation-maximisation algorithm, souvent abrégé EM), proposé par Dempster et al. (1977), est une classe d'algorithmes qui permettent de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables (*Wikipedia*). 33, 36, 39, 44, 111, 186
GMM	Modèle de mélanges de gaussiennes ou *Gaussian Mixture Model* (GMM). 52, 103, 105, 107–109, 115, 124, 227
HDP	Processus Hiérarchique de Dirichlet ou *Hierarchical Dirichlet Process*. 77
IA	L'intelligence artificielle est la "recherche de moyens susceptibles de doter les systèmes informatiques de capacités intellectuelles comparables à celles des êtres humains" (*Wikipedia*). 18
IDF	Fréquence inverse de documents ou *Inverse Document Frequency*. 12, 17, 30, 31, 37, 50, 51, 53, 54, 57–61, 64–66, 70, 96–98, 141–144, 157–159, 161, 195, 196
JFA	Analyse Factorielle Jointe ou Joint Factor Analysis. 103, 106–108

KNN En intelligence artificielle, la méthode des k plus proches voisins est une méthode d'apprentissage supervisé. En abrégé k-NN ou KNN, de l'anglais k-nearest neighbor. Dans ce cadre, on dispose d'une base de données d'apprentissage constituée de N couples "entrée-sortie". Pour estimer la sortie associée à une nouvelle entrée x, la méthode des k plus proches voisins consiste à prendre en compte (de façon identique) les k échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée x, selon une distance à définir. Par exemple, dans un problème de classification, on retiendra la classe la plus représentée parmi les k sorties associées aux k entrées les plus proches de la nouvelle entrée x (*Wikipedia*). 142, 144, 145

LDA Allocation Latente de Dirichlet ou *Latent Dirichlet Allocation*. 35, 37–45, 49, 52–55, 57–66, 68–70, 77–82, 85, 88, 91, 92, 95–99, 102, 103, 109–112, 116–119, 123, 124, 127, 133, 138, 149, 150, 184, 195, 196, 199

LDLM Modèle de Langage Latent de Dirichlet (MLLD) ou *latent Dirichlet language model*. 43, 44

LSA Analyse Sémantique latente ou *Latent Semantic Analysis*. 32–35, 41, 163, 167, 171, 184

LSI Indexation Sémantique latente ou *Latent Semantic Indexation*. 165, 167

MCMC Markov Chain Monte-Carlo. 39, 40

PLDA Probabilistic Linear Discriminant Analysis. 108

PLSA Analyse Sémantique latente Probabiliste ou *Probabilistic Latent Semantic Analysis*. 34–39, 41, 43, 44, 171, 173, 174, 196

QR Question-Réponse ou *Question-Answering* (QA). 16

RAP Reconnaissance Automatique de la Parole. 36, 44, 49, 58

RATP La Régie autonome des transports parisiens (RATP) est un établissement public à caractère industriel et commercial (EPIC) assurant l'exploitation d'une partie des transports en commun de Paris et de sa banlieue. Elle exploite les seize lignes du métro de Paris, six des lignes du tramway d'Île-de-France (T1, T2, T3a, T3b, T5, et T7. La ligne T4 est exploitée par la SNCF), une partie des lignes de bus d'Île-de-France, et une partie des lignes A et B du réseau express régional d'Île-de-France (RER). En région parisienne, elle transporte environ 3 milliards de passagers par an (2010) (*Wikipedia*). 138, 185

RI La recherche d'information (RI) est le domaine qui étudie la manière de retrouver des informations dans un corpus. Celui-ci est composé de documents d'une ou plusieurs bases de données, qui sont décrits par un contenu ou les métadonnées associées. Les bases de données peuvent être relationnelles ou non structurées, telles celles mises en réseau par des liens hypertextes comme dans le World Wide Web, l'Internet et les Intranets. Le contenu des documents peut être du texte, des sons, des images ou des données (*Wikipedia*). 16–19, 30, 32, 33, 41–43, 48, 50, 68

RP Position Relative ou *Relative Position*. 70

SGMM Modèle de sous-espace de mélanges de gaussiennes ou *Subspace Gaussian Mixture Model* (SGMM). 107
SRAP Système de Reconnaissance Automatique de la Parole. 44, 49–51, 133, 142
STM Message court ou *Short Text Message*. 67–70
SVD Décomposition en Valeurs Singulières ou Singular Value Decomposition. 63, 77
SVM Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais Support Vector Machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les SVM sont une généralisation des classifieurs linéaires (*Wikipedia*). 50, 52, 53, 55, 58–62, 66, 78, 81, 82, 84, 85, 88–91, 96–98, 105, 195, 199

TALN Traitement Automatique du Langage naturel ou Natural Language Processing (NLP) est une discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain. Ainsi, le TAL ou TALN est parfois nommé ingénierie linguistique (*Wikipedia*). 16, 18, 19
TCLDLM *Topic Cache Latent Dirichlet language model* (TCLDLM). 44
TEM Taux d'Erreur-Mot (TEM) ou *Word Error Rate*. 43, 44, 51, 58, 61, 64, 65, 92, 95, 97, 99, 186
TF Fréquence de mots ou *Term Frequency*. 12, 30, 31, 37, 50, 51, 53, 54, 57–61, 64–66, 70, 96–98, 141–144, 157–159, 161, 195, 196
TM espace de thèmes ou *topic model*. 32
TMAN Transcription manuelle d'un document. 53, 58–66, 95, 97, 98, 119, 120, 122, 124, 143–145, 196, 199
TRAP Transcription provenant d'un Système de Reconnaissance Automatique de la Parole. 53, 58–62, 64–66, 95, 97, 98, 107, 119, 121–124, 143–145, 196, 199
TREC Text REtrieval Conference. 43

UBM Le modèle du monde (appelé Universal Background Model (UBM)) est typiquement un GMM (GMM) entraîné sur un l'ensemble des données d'apprentissage en utilisant le critère de vraisemblance maximum. 103, 107–109, 124

VB Variational Bayesian. 39, 43
VBI Variational Bayesian Inference. 39
VL Vérification du Locuteur. 104, 105

WSJ *The Wall Street Journal* est un quotidien national américain qui traite de l'actualité économique et financière, créé à New York par Dow Jones and Company (*Wikipedia*). 44
WWW Le World Wide Web (WWW), littéralement la "toile (d'araignée) mondiale", communément appelé le Web, et parfois la Toile, est un système hypertexte public fonctionnant sur Internet (*Wikipedia*). 16

Bibliographie personnelle

Journal International

[Morchid et al., 2014] Morchid, M., Dufour, R., Bousquet, P.-M., Linarès, G., and Torres-Moreno, J.-M. (2014). Feature selection using principal component analysis for massive retweet detection. *Pattern recognition letters, Edition Elsevier*.

Conférences Internationales

[Bouallegue et al., 2014a] Bouallegue, M., Morchid, M., Dufour, R., Driss, M., Linarès, G., and De Mori, R. (2014a). Factor analysis based semantic variability compensation for automatic conversation representation. In *Conference of the International Speech Communication Association (INTERSPEECH) 2014*. ISCA.

[Bouallegue et al., 2014b] Bouallegue, M., Morchid, M., Dufour, R., Driss, M., Linarès, G., and De Mori, R. (2014b). Subspace gaussian mixture models for dialogues classification. In *Conference of the International Speech Communication Association (INTERSPEECH) 2014*. ISCA.

[Loni et al., 2014] Loni, B., Hare, J., Georgescu, M., Riegler, M., Zhu, X., Morchid, M., Dufour, R., and Larson, M. (2014). Getting by with a little help from the crowd: Practical approaches to social image labeling. In *the International Workshop on Crowdsourcing for Multimedia (CrowdMM) 2014*. ACM.

[Morchid et al., 2014a] Morchid, M., Bouallegue, M., Dufour, R., Linarès, G., Matrouf, D., and De Mori, R. (2014a). An i-vector based approach to compact multi-granularity topic spaces representation of textual documents. In *the Conference of Empirical Methods on Natural Lnguage Processing (EMNLP) 2014*. SIGDAT.

[Morchid et al., 2014b] Morchid, M., Bouallegue, M., Dufour, R., Linarès, G., Matrouf, D., and De Mori, R. (2014b). I-vector based representation of highly imperfect automatic transcriptions. In *Conference of the International Speech Communication Association (INTERSPEECH) 2014*. ISCA.

[Morchid et al., 2014c] Morchid, M., Dufour, R., Bouallegue, M., and Linarès, G. (2014c). Author-topic based representation of call-center conversations. In *International Spoken Language Technology Workshop (SLT) 2014*. IEEE.

[Morchid et al., 2014d] Morchid, M., Dufour, R., Bouallegue, M., Linarès, G., and De Mori, R. (2014d). Theme identification in human-human conversations with features from specific speaker type hidden spaces. In *Conference of the International Speech Communication Association (INTERSPEECH) 2014*. ISCA.

[Morchid et al., 2014e] Morchid, M., Dufour, R., Bousquet, P.-M., Bouallegue, M., Linarès, G., and De Mori, R. (2014e). Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule. In *International Conference on Acoustic, Speech and Signal Processing (ICASSP) 2014*. IEEE.

[Morchid et al., 2013a] Morchid, M., Dufour, R., and Linarès, G. (2013a). Event detection from image hosting services by slightly-supervised multi-span context models. In *International Workshop on Content-Based Multimedia Indexing (CBMI) 2013*. IEEE.

[Morchid et al., 2013b] Morchid, M., Dufour, R., and Linarès, G. (2013b). Thematic representation of short text messages with latent topics: Application in the twitter context. In *Conference of the Pacific Association for Computational Linguistics (PACLING) 2013*.

[Morchid et al., 2014f] Morchid, M., Dufour, R., and Linarès, G. (2014f). A combined thematic and acoustic approach for a music recommendation service for tv commercials. In *International Society for Music Information Retrieval Conference (ISMIR) 2014*.

[Morchid et al., 2014g] Morchid, M., Dufour, R., and Linarès, G. (2014g). A lda-based topic classification approach from highly imperfect automatic transcriptions. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC) 2014*.

[Morchid et al., 2014h] Morchid, M., Dufour, R., Niaz, U., Bouvier, F., de Groc, C., Linarès, G., Meiraldo, B., and Peralta, B. (2014h). Sumacc project's corpus: A topic-based query extention approach to retrieve multimedia documents. In *Proceedings of the 17th International Conference on Text, Speech and Dialogue (TSD) 2014*. ISCA.

[Morchid and Linarès, 2013] Morchid, M. and Linarès, G. (2013). Lda-based method for automatic tagging of youtube videos. In *International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS) 2013*. IEEE.

[Morchid et al., 2014i] Morchid, M., Linarès, G., and Dufour, R. (2014i). Characterizing and predicting bursty events: The buzz case study on twitter. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC) 2014*.

[Morchid et al., 2013c] Morchid, M., Linarès, G., El-Beze, M., and De Mori, R. (2013c). Theme identification in telephone service conversations using quaternions of speech features. In *Conference of the International Speech Communication Association (INTERSPEECH) 2013*. ISCA.

Conférences Nationales

[Morchid et al., 2013] Morchid, M., Dufour, R., and Linarès, G. (2013). Combinaison de thèmes latents pour la contextualisation de tweets. *Atelier Contextualisation de Messages Courts (EGC) 2013*.

[Morchid et al., 2014] Morchid, M., Dufour, R., Linarès, G., and De Mori, R. (2014). Classification de transcriptions automatiques imparfaites : Doit-on adapter le calcul du taux d'erreur-mot ? In *Proceedings Journées d'étude de la parole (JEP) 2014*.

[Morchid and Linarès, 2012] Morchid, M. and Linarès, G. (2012). Extraction de mots-clefs dans des vidéos web par analyse latente de dirichlet. In *Proc. Journées d'étude de la parole (JEP) 2012*.

[Morchid and Linarès, 2013] Morchid, M. and Linarès, G. (2013). Prédiction des buzz sur twitter. In *COnfrence en Recherche d'Information et Applications (CORIA) 2013*.

Ateliers

[Bost et al., 2013] Bost, X., Brunetti, I., Cabrera-Diego, L. A., Cossu, J.-V., Linhares, A., Morchid, M., Torres-Moreno, J.-M., El-Bèze, M., and Richard, D. (2013). Systèmes du lia à deft'13. In *Actes du neuvième DÉfi Fouille de Textes (DEFT) 2013*.

[Bouallegue et al., 2013] Bouallegue, M., Senay, G., Morchid, M., Matrouf, D., Linarès, G., and Dufour, R. (2013). Lia @ mediaeval 2013 spoken web search task: An i-vector based approach. In *(MediaEval) 2013*.

[Cossu et al., 2013] Cossu, J.-V., Bigot, B., Bonnefoy, L., Morchid, M., Bost, X., Senay, G., Dufour, R., Bouvier, V., Torres-Moreno, J.-M., and El-Bze, M. (2013). Lia@replab 2013. In *An evaluation campaign for Online Reputation Management Systems (CLEF) 2013*.

[Morchid et al., 2013a] Morchid, M., Dufour, R., Bouallegue, M., Linarès, G., and Matrouf, D. (2013a). Lia @ mediaeval 2013 crowdsourcing task: Metadata or not metadata? that is a fashion question. In *(MediaEval) 2013*.

[Morchid et al., 2013b] Morchid, M., Dufour, R., Bouallegue, M., Linarès, G., and Matrouf, D. (2013b). Lia @ mediaeval 2013 musiclef task: A combined thematic and acoustic approach. In *(MediaEval) 2013*.

[Morchid et al., 2014] Morchid, M., Huet, S., and Richard, D. (2014). A topic-based approach for post-processing correction of automatic translations. In *International Workshop on Spoken Language Translation (IWSLT) 2014*.

[Morchid and Linarès, 2011] Morchid, M. and Linarès, G. (2011). Mediaeval benchmark: Social event detection using lda and external resources. In *(MediaEval) 2011*.

[Morchid and Linarès, 2012a] Morchid, M. and Linarès, G. (2012a). Bursty event prediction by content-based approach. In *Third Workshop of the Brazilian Institute for Web Science Research 2012*.

[Morchid and Linarès, 2012b] Morchid, M. and Linarès, G. (2012b). Inex 2012 benchmark a semantic space for tweets contextualization. In *Workshop of the Initiative for the Evaluation of XML Retrieval (CLEF) 2012*.

REPRESENTATIONS ROBUSTES DE DOCUMENTS BRUITÉS DANS DES ESPACES HOMOGÈNES

Résumé

En recherche d'information, les documents sont le plus souvent considérés comme des "sacs-de-mots". Ce modèle ne tient pas compte de la structure temporelle du document et est sensible aux bruits qui peuvent altérer la forme lexicale. Ces bruits pouvent être produits par différentes sources : forme peu contrôlée de messages issus du Web, messages vocaux dont la transcription automatique contient des erreurs… Le travail présenté dans cette thèse s'intéresse au problème de la représentation de documents issus de sources bruitées. Dans un premier temps, nous comparons une représentation classique, utilisant la fréquence des mots, à une représentation de haut niveau s'appuyant sur un espace de thèmes. Le problème majeur d'une telle représentation est qu'elle est fondée sur un espace de thèmes dont les paramètres sont choisis empiriquement. Nous décrivons ensuite une représentation originale qui s'appuie sur des espaces multiples pour résoudre trois problèmes majeurs : la proximité des sujets traités dans le document, le choix difficile des paramètres du modèle de thèmes ainsi que la robustesse de la représentation. Partant de l'idée qu'une seule représentation des contenus ne peut pas capturer l'ensemble des informations utiles, nous proposons d'augmenter le nombre de vues sur un même document. Cette multiplication des vues est efficace mais elle a l'inconvénient d'être très volumineuse, redondante et de contenir une variabilité additionnelle liée à la diversité des vues. Nous proposons de traiter ces problèmes avec une méthode basée sur l'analyse factorielle pour fusionner les vues multiples et obtenir une nouvelle représentation robuste, de dimension réduite, ne contenant que la partie "utile" du document tout en réduisant les variabilités "parasites". Enfin, lors de l'élaboration des espaces de thèmes, le document reste considéré comme un "sac-de-mots" alors que plusieurs études montrent que la position d'un terme au sein du document est importante. Une représentation basée sur les *quaternions*, tenant compte de cette structure temporelle du document est finalement proposée.

Mots clés : *Représentation robuste, document bruité, allocation latente de Dirichlet, représentation multi-vues, analyse factorielle, quaternion.*
Keywords : *Robust representation, noisy document, latent Dirichlet allocation, multi-views representation, factor analysis, quaternion.*